Group Sequential and Adaptive Methods – Topics with Applications to Clinical Trials

submitted by

Carl Fredrik Öhrn

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

February 2011

Summary

This thesis deals with sequential and adaptive methods for clinical trials, and how such methods can be used to achieve efficient clinical trial designs. The efficiency gains that can be achieved through non-adaptive group sequential methods are well established, while the newer adaptive methods seek to combine the best of the classical group sequential framework with an approach that gives increased flexibility.

Our results show that the adaptive methods can provide some additional efficiency, as well as increased possibilities to respond to new internal and external information. Care is however needed when applying adaptive methods. While sub-optimal rules for adaptation can lead to inefficiencies, the logistical challenges can also be considerable. Efficient non-adaptive group sequential designs are often easier to implement in practice, and have for the cases we have considered been quite competitive in terms of efficiency.

The four problems that are presented in this thesis are very relevant to how clinical trials are run in practice. The solutions that we present are either new approaches to problems that have not previously been solved, or methods that are more efficient than the ones currently available in the literature. Several challenging optimisation problems are solved through numerical computations. The optimal designs that are achieved can be used to benchmark new methods proposed in this thesis as well as methods available in the statistical literature.

The problem that is solved in Chapter 5 can be viewed as a natural extension to the other problems. It brings together methods that we have used to the design of individual trials, to solve the more complex problem of designing a sequence of trials that are the core part of a clinical development program. The expected utility that is maximised is motivated by how the development of new medicines works in practice.

Acknowledgments

It is a pleasure to thank everyone who has assisted me while writing this thesis. Firstly, I would like to thank my supervisor, Prof. Christopher Jennison, for teaching me about a fascinating subject and for giving me invaluable support and guidance throughout my PhD studies. I have also benefited greatly from the supervision of Dr. Carl-Fredrik Burman, who has provided insightful advice, stimulating discussions and valuable feedback on the manuscript. I would also like to thank Dr. Simon Shaw for providing useful comments after reviewing my progress reports and Dr. Christian Sonesson for supervision at the beginning of the PhD studies. I must also mention Dr. Ziad Taib, who several years ago gave me the opportunity to write a Master of Science Thesis in biostatistics and more recently organised an extended visit for Prof. Jennison in Sweden.

I am indebted to Dr. Carl-Fredrik Burman and Dr. Christian Sonesson for initiating the PhD project, to Anders Carlquist, who as Section Director of Biostatistics supported the idea and to AstraZeneca for making the PhD studies possible. I would also like to take this opportunity to thank everyone at the Biostatistics group at AstraZeneca and at the Department of Mathematical Sciences, University of Bath, for providing a friendly and inspiring work environment. Special thanks must go to my line managers at AstraZeneca, who where always supportive of the PhD project.

Finally, I could not have completed this PhD thesis without my family and friends. In particular, I would like to express my gratitude to my parents. You have always encouraged me to study, even though I chose a field of research quite different from yours. Last but not least, I would like to thank my wonderful wife Marlo for sharing this journey with me. Thank you Marlo for all your love, patience, understanding and support during the last four years.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Glossary

cdf	cumulative distribution function
CP_θ	conditional power at effect size of θ
CRP	conditional rejection probability
$\text{CRP}_{\theta=0}$	conditional rejection probability calculated for $\theta = 0$
EMEA	European Medicines Agency
FDA	Food and Drug Administration
H_0	null hypothesis
H_1	alternative hypothesis
ICH	International Conference of Harmonisation
$\mathcal{I}_{fix}$	Information in fixed sample trial
$\mathcal{I}_k$	information at analysis k of a group sequential design
$\mathcal{I}_{max}$	maximum information level for a group sequential design
K	number of analyses in a group sequential design
$N(\theta, \sigma^2)$	normal distribution with mean θ and variance σ^2
n_{max}	maximum per-group sample size for a group sequential design
P_θ	probability given effect size of θ
pdf	probability density function
$p_{kb}(z_k)$	conditional probability of not having crossed the upper boundary, given $Z_k = z_k$
$p_{kc}(z_k)$	conditional probability of not having crossed the lower or upper boundary, given $Z_k = z_k$
PP	predictive power
r	correlation in bivariate normal model for mean of phase II and phase III endpoints
SPRT	sequential probability ratio test
WMA	World Medical Association
$z_{1-\alpha}$	$100 \times (1-\alpha)\%$ quantile of a standard normal distribution
α	type I error probability
β	type II error probability
δ	effect size at which power is set in two-decision problem
δ_N	non-inferiority margin
δ_S	effect size at which power for superiority is set in three-decision problem
Φ	standard normal cumulative distribution function
ϕ	standard normal probability density function
θ	mean of the treatment effect
$\hat{\theta}$	maximum likelihood estimate of θ
θ_2	mean of the treatment effect in phase II
θ_3	mean of the treatment effect in phase III
$\pi(\theta)$	prior distribution for θ
$\pi^{(k)}(\theta \mid z_k)$	posterior distribution for θ given $Z_k = z_k$

CHAPTER 1

Introduction

1.1 Development of new medicines

Over the past decades, great advances have been made in medicine. There are, however, many conditions that remain without satisfactory treatment options. Moreover, there may be room for further improvements in areas where important progress has already been achieved. New drugs have for example been successful in treating cardiovascular risk factors like elevated cholesterol and hypertension, with associated reductions in cardiovascular events such as myocardial infarction and stroke. Despite this progress, cardiovascular disease remains one of the most important causes of death around the world. In other areas, the drugs that are available can slow the progression of a disease somewhat, without actually providing a cure. Further improvement through new medicines may then still be possible, in terms of reduced mortality or improved quality of life.

There are clearly many areas of unmet medical need, where new medicines could provide important improvements to the life of patients. Before a new medicine can be made publicly available, it must be shown to be safe and efficacious in the patient population where it will be used. Randomised clinical trials, where patients are randomly assigned to one of several treatment groups, are considered to be the most reliable way to evaluate a new experimental drug. They make it possible to make causal inference about the treatment effect, and assess the benefits and risks of different treatment regimens. Such risk/benefit assessments form the basis for regulatory authorities when deciding whether to approve a new drug.

Whenever possible from a logistical perspective, it is desirable that randomised clinical trials are double-blind. In a double-blind trial, neither the patient nor the investigator knows to which treatment group the patient has been randomised. The intention is to remove the possibility for the investigator to approach patients

differently, depending on to which treatment group the patients have been randomised. Making sure that a trial is double-blind is thought to reduce bias and increase the scientific credibility of the results.

Conducting a double-blind randomised clinical trial can present difficult ethical issues. In some situations it may for example be considered unethical to randomise patients to placebo, particularly for conditions where a safe and efficacious treatment is already available for public use. As stated in the declaration of Helsinki, it is fundamental that patients make informed decisions about whether they wish to participate in a clinical trial, both initially and throughout the duration of the trial (World Medical Association, 2008). This process is referred to as patients giving their informed consent to participating in the trial. Patients providing their informed consent is one of the key principles of good clinical practice, which should be followed for a clinical trial to be credible. The International Conference for Harmonisation (ICH) Guideline E6 aims to protect the ethical and scientific quality of clinical trials and is a good source for further information about such matters (ICH, 1996).

Clinical trials have traditionally been divided into four phases. The characteristics of each phase can differ between therapeutic areas, but can broadly be outlined as follows:

- Phase I clinical trials aim at making an initial assessment of the safety, tolerability, pharmacokinetics, and pharmacodynamics of a drug. They often include a small group of healthy volunteers, but in some therapeutic areas such as oncology, patients are involved already in this early phase.

- Phase II clinical trials typically include more subjects than phase I trials, but are still of moderate size. An important part of phase II is to evaluate both efficacy and safety aspects of the drug at different dose levels. Based on phase II data, a decision about whether to progress to phase III has to be made. At the end of phase II, it is also decided which dose(s) to bring forward to phase III.

- The objective of phase III clinical trials is to demonstrate that an experimental drug is efficacious and safe. This is the phase where the statistical framework of hypothesis testing is most rigorously applied. If the programme of phase III trials is successful, submissions for regulatory approval are made to regulatory agencies around the world. Phase III clinical trials often include a large number of patients recruited at many centres in different countries. They are costly to run and may have a long treatment duration and follow-up.

- Phase IV clinical trials are conducted after a drug has received regulatory approval. Some Phase IV trials are required by regulatory authorities and may be referred to as post marketing surveillance trials. In other cases the focus may

be on life cycle management activities, which aim to broaden the indication of the drug and increase sales.

The main focus of this thesis will be on sequential and adaptive methods used in phase III clinical trials. In addition, we shall study the joint planning of phase II and phase III trials in Chapter 5. In a phase III clinical trial, the primary objective is often to confirm that the experimental drug is efficacious compared to control. Suppose that the true treatment effect is θ_E for the experimental drug and θ_C for the control, and that a positive value of θ_E or θ_C implies that the treatment has been beneficial to the patient. The treatment effect $\theta = \theta_E - \theta_C$ can then be assessed in a statistical hypothesis test framework, where the null hypothesis $H_0 : \theta \leq 0$, is tested against the alternative hypothesis $H_1 : \theta > 0$. It is a regulatory requirement to control the type I error, the probability of falsely rejecting the null hypothesis, at some pre-specified level α. Regulators typically require $\alpha \leq 0.025$ for one-sided tests, which corresponds to allowing $\alpha \leq 0.05$ for two-sided tests. It is also desirable to make sure that the power at a certain value of the treatment effect, $\theta = \delta$ say, is at least $1 - \beta$. The choices of β and δ are not as strictly controlled by regulators. The type II error β can be viewed as the sponsor's risk of an efficacious treatment not achieving a statistically significant result. Common choices for β are $\beta = 0.1$ and $\beta = 0.2$.

In this thesis we are concerned with methods for making the design of clinical trials more efficient. Given α, β, δ and some additional assumptions, we can calculate the number of patients needed to satisfy the power requirement in a fixed sample trial. Unfortunately, this number is often rather high, which makes phase III trials very expensive. The increased costs of drug development over the last decade are discussed by DiMasi et al. (2003). They estimate the average cost spent prior to approval of a new drug to be approximately US$ 800 million. This underlines the importance of efficient clinical trials, so that more treatments can be tested, approved and made available to patients. One way to make clinical trials more efficient is to apply group sequential methods, which will be discussed in the next section.

1.2 Group sequential designs

1.2.1 Background

In the early literature about sequential methods, data were typically assumed to be monitored continuously, i.e. for every patient. There was an early focus on applications in quality control, which gained importance during World War II when it was essential to make sure that ammunition was of appropriate quality. For testing the simple null hypothesis $\theta = \theta_0$ against the alternative $\theta = \theta_1$, Wald (1945, 1947) introduced the sequential probability ratio test (SPRT). In the SPRT, the likelihood ratio of the

accumulated data, under $\theta = \theta_1$ and $\theta = \theta_0$, is compared to critical values A and B. The test is stopped to reject H_0 if the likelihood ratio is larger than A and to accept H_0 if the likelihood ratio is smaller than B. A and B are chosen to make sure that the test has approximately type I error probability α and power $1-\beta$ at $\theta = \theta_1$. Wald and Wolfowitz (1948) proved that the SPRT under certain assumptions is optimal, in the sense that it minimises the expected sample size. One drawback with the SPRT is that there is no maximal sample size at which sampling is guaranteed to stop. Truncated versions of the test, with a maximal sample size, have been proposed to deal with this issue.

1.2.2 Two-sided group sequential tests with fixed group sizes

In clinical trial applications, it would often be too logistically challenging to monitor data continuously. Continuous monitoring would be particularly inconvenient in large scale trials. If thousands of patients are distributed across hundreds of centres around the world, it would not be possible to perform a new interim analysis for every patient. Pocock (1977) therefore suggested analysing data after certain cumulative sample sizes had been accrued. Performing repeated significance tests at significance level α will however lead to an overall type I error probability that is larger than α. Hence, the nominal significance level at each interim analysis is adjusted downwards in Pocock's test, to account for the multiple looks and get an overall type I error probability of α. The test stops to reject $H_0 : \theta = 0$ if

$$|Z_k| \geq b_k, \quad k = 1, \ldots, K,$$

where Z_k is usual standardised statistic at analysis k. The critical values b_k are in Pocock's test constant on the standardised Z statistic scale, for $k = 1, \ldots, K$. The critical values can be written as

$$b_k = C_{\text{P}}(K, \alpha),$$

where C_{P} depends on the total number of analyses K and the type I error probability α.

The test of Pocock (1977) is an example of a two-sided test of the null hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta \neq 0$. The critical values are set so that $P_{\theta=0}(\text{Reject } H_0) = \alpha$, while the power requirement is that $P_{\theta=\pm\delta}(\text{Reject } H_0) = 1-\beta$. O'Brien and Fleming (1979) proposed a two-sided group sequential test with much wider boundaries at earlier interim analyses and narrower boundaries later. The critical values in O'Brien and Fleming's test can be written as

$$b_k \quad = \quad C_{\text{OF}}(K, \alpha)\sqrt{\frac{K}{k}}, \text{ for } k = 1, \ldots, K,$$

where C_{OF} is the critical value at the final analysis and depends on the total number of analyses K and the type I error α. Wang and Tsiatis (1987) defined a very general family of tests where the parameter Δ decides the amount of early stopping. This class of designs contains the Pocock design the O'Brien and Fleming design as special cases, for particular values of Δ. In the tests of Wang and Tsiatis (1987), the critical values can be written as

$$b_k \;=\; C_{\mathrm{WT}}(K,\alpha,\Delta)\left(\frac{k}{K}\right)^{\Delta-1/2},\ \text{for } k=1,\ldots,K,$$

where C_{WT} is the critical value at the final analysis and depends on K, α and Δ. We see that $\Delta = 0.5$ corresponds to the test of Pocock (1977) and $\Delta = 0$ to the test of O'Brien and Fleming (1979). The correct choice of C_{WT}, for a given combination of K, α and Δ, is tabulated in Jennison and Turnbull (2000, Chapter 2). This information is also available in the software package East-5 (2007).

1.2.3 Sequential distribution theory

Suppose that the observations in treatment groups A and B are independent and normally distributed with $X_{Ai} \sim N(\mu_A,\sigma^2)$ and $X_{Bi} \sim N(\mu_B,\sigma^2)$, where the common variance σ^2 is known. Assume further that the two treatments are compared in a group sequential trial where the primary objective is to make inference about the parameter $\theta{=}\mu_B - \mu_A$. At each interim analysis we can calculate $\hat{\theta}_k$, the maximum likelihood estimate for θ at analysis k, according to

$$\hat{\theta}_k = \sum_{i=1}^{n_k}(X_{Bi} - X_{Ai})/n_k$$

where n_k is the cumulative per-group sample size. Let $\mathcal{I}_k$ denote the Fisher information for θ at analysis k, defined as

$$\mathcal{I}_k = \frac{n_k}{2\sigma^2}.$$

The maximum likelihood estimate $\hat{\theta}_k$ then follows a normal distribution according to

$$\hat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1}).$$

If $Z_k = \hat{\theta}_k\sqrt{\mathcal{I}_k}$ denotes the standardised statistic at analysis k, the sequence of statistics $Z_1,\ldots,Z_K$ follow what Jennison and Turnbull (2000, Chapter 3) refer to as the canonical joint distribution. The statistics $Z_1,\ldots,Z_K$ have the canonical joint

distribution with information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ for θ if

$$\begin{aligned} &(Z_1, \ldots, Z_K) \text{ is multivariate normal,} \\ &E(Z_k) = \theta\sqrt{\mathcal{I}_k}, \quad k = 1, \ldots, K, \text{ and} \\ &\mathrm{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}, \quad 1 \leq k_1 \leq k_2 \leq K. \end{aligned} \tag{1.1}$$

The sequence $Z_1, \ldots, Z_K$ is a Markov sequence, and this is important for the numerical integration methods used to calculate properties of group sequential tests. We have so far assumed data with normal response and known variance, but note that (1.1) holds at least approximately in many other situations, for example survival data or normally distributed data adjusted for baseline covariates. We refer to Jennison and Turnbull (2000) for a comprehensive acccount of how to construct group sequential tests for a wide range of response distributions.

1.2.4 One-sided group sequential tests with fixed group sizes

The designs in Section 1.2.2 were proposed for two-sided tests, but in clinical trials we are often interested in one-sided testing problems. If θ represents the treatment benefit of the experimental drug compared to placebo, it may for example be of interest to establish the presence of a positive treatment effect by rejecting $H_0 : \theta \leq 0$. The distribution theory in Section 1.2.3 is applicable also for group sequential tests. Hence, such tests can be accommodated within a framework similar to that used for the two-sided tests in Section 1.2.2. In one-sided tests, there is an upper boundary for rejecting H_0, and there may also be a lower boundary for accepting H_0. Stopping early to accept H_0 is often referred to as stopping for futility, as the trial is deemed futile and it is unlikely that the results that were hoped for at the outset can be achieved. One-sided group sequential tests with futility boundaries are considered in some detail in Chapter 4, while one-sided tests without futility boundaries are used in Chapter 3. In the context of two-sided tests, Gould and Pecore (1982) proposed including boundaries for early stopping to accept the null hypothesis. A variation of this type of design is given in Chapter 2, where some of the error probabilities have different roles than in the designs of Gould and Pecore (1982).

Let us now consider how one-sided group sequential tests can be set up when there is no futility boundary. We consider the clinical trial described in the beginning of Section 1.2.3, where the goal is to make inference about θ by testing the null hypothesis $H_0 : \theta \leq 0$ against the alternative hypothesis $H_1 : \theta > 0$, at significance level $\alpha = 0.025$ and power $1 - \beta$ at $\theta = \delta$. At each interim analysis k, we stop to reject H_0 if

$$Z_k = \hat{\theta}_k\sqrt{\mathcal{I}_k} \geq b_k,$$

where the critical values b_k are chosen so that the overall type I error probability equals α. To find critical values $b_1, \ldots, b_K$ that give a test with type I error probability α, we can make use of the upper boundary of the two-sided tests of Pocock (1977), O'Brien and Fleming (1979) or Wang and Tsiatis (1987), which are described in Section 1.2.2. Suppose that one of these tests has two-sided type I error probability 2α. The probability of first crossing one of the boundaries of the two-sided tests and then the other is usually negligible. If this small probability is ignored, a one-sided group sequential test, with the upper boundary of any of these two-sided tests but no lower boundary, will have type I error probability α.

When planning the design, it is of interest to know $\mathcal{I}_{max}$, the amount of information needed at the final analysis to satisfy the power requirements. Suppose that it has been decided to use the group sequential boundary of O'Brien and Fleming (1979). By assuming a particular form for the information sequence, for example equally spaced analyses, $\mathcal{I}_{max}$ can be found by searching for the maximum information that gives power $1 - \beta$ at $\theta = \delta$. The relationship between $\mathcal{I}_{max}$ and the corresponding fixed sample size $\mathcal{I}_{fix}$ can be expressed as

$$\mathcal{I}_{max} \;=\; R\,\mathcal{I}_{fix},$$

where

$$\mathcal{I}_{fix} \;=\; \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta^2}$$

and R is an inflation factor. For any group sequential design with early stopping, we will have $R > 1$, and hence, $\mathcal{I}_{max} > \mathcal{I}_{fix}$. The expected information of the group sequential test will however typically be lower than for the corresponding fixed sample test.

1.2.5 Error spending designs

The tests described in Sections 1.2.2 and 1.2.4 all assume fixed information sequences. If the observed information sequence differs from the one used to derive the critical values, prior to the start of the trial, the type I error probability will no longer be equal to α. Lan and DeMets (1983) proposed error spending designs as a way of dealing with unpredictable information sequences. In error spending designs, the cumulative type I error probability is specified as a function of the observed information. The critical values are derived by using the fact that the sequence of test statistics $Z_1, \ldots, Z_K$ follow the joint canonical distribution in (1.1). Let us now consider how to find the critical value at analysis k, of a one-sided error spending design without futility boundary. Define a non-decreasing type I error spending function $f(\mathcal{I})$, which satisfies $f(0) = 0$ and $f(\mathcal{I}) = \alpha$, for $\mathcal{I} \geq \mathcal{I}_{max}$. At the first interim analysis, the critical value b_1 is found by solving $P_{\theta=0}(Z_1 \geq b_1) = f(\mathcal{I}_1)$. The critical value b_k at analysis k is calculated as

the solution to

$$P_{\theta=0}(Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k \geq b_k) = f(\mathcal{I}_k) - f(\mathcal{I}_{k-1}). \tag{1.2}$$

At the planning stage, the inflation factor R and the maximal information level $\mathcal{I}_{max}$ needed to meet the power requirement are found in the same way as for group sequential designs with fixed information sequences. By construction, the error spending method gives type I error probability of exactly α. Furthermore, Jennison and Turnbull (2000, Chapter 7) show that for moderate changes to the planned information sequences, the attained type II error probability is close to its target β. A key assumption is that the information levels at which future interim analyses take place must not be chosen based on the observed treatment effect. Proschan et al. (1992) have shown that violations of this assumption can lead to serious inflation of the type I error rate. The critical values for two-sided error spending tests and one-sided error spending tests with futility boundaries can be found using a similar approach. In Chapter 2 we shall also introduce error spending designs for group sequential designs that simultaneously test superiority and non-inferiority.

Probabilities like (1.2) can be calculated through numerical integration. The calculations in this thesis are based on the methods described by Jennison and Turnbull (2000, Chapter 19). Jennison and Turnbull suggest using a grid of $m = 189$ points to integrate a normal density over the real line. Two thirds of the grid points are within $\pm$ three standard deviations of the mean. The method makes use of the thin tails of the normal distribution, using a logarithmic spacing for the final third of the grid points. Chandler and Graham (1988) show that $O(m^{-4})$ convergence is obtained, when integrating a normal density with this type of grid, using Simpson's rule.

1.2.6 Inference on termination

At termination of a group sequential design, it is desirable to make inference about the treatment effect θ and present an unbiased point estimate, confidence interval and p-value. Because of the stopping rule that has been applied, this is not as straightforward as for fixed sample trials. If a group sequential design is stopped early for benefit, it can for example be shown that the standard maximum likelihood estimate of θ at termination is biased. Further details about inference on termination of a group sequential test are provided by Proschan et al. (2006, Chapter 7) and Jennison and Turnbull (2000, Chapter 8).

Inference on termination of a group sequential design can be further complicated by the issue of overrunning. Overrunning is a concept that has typically been overlooked in the literature about group sequential designs. It can occur when there is a delay in

time between when a patient is randomised and when the response is observed. When a decision is made to stop a trial early after an efficacy boundary has been crossed, there are likely to be additional patients who have recently been randomised but were not included in the interim analysis. This makes inference on termination even more complex, as it would be desirable to include data for these additional patients in the final analysis. Hampson (2009) gives a very thorough treatment of how to deal with these issues, including methods for how to derive confidence intervals and p-values.

1.2.7 Benefits of group sequential designs

As the statistical aspects of a group sequential design are clearly more complex than for a fixed sample trial, it is reasonable to ask what the benefits with applying group sequential methods are. Jennison and Turnbull (2000, Chapter 1) divide the benefits of group sequential methods into ethical, administrative and economical.

- Throughout a clinical trial, there is an ethical need to monitor the safety of the patients in all treatment arms. This can for example be done through group sequential monitoring. If a group sequential trial is stopped early for a positive effect, fewer patients are exposed to an inferior treatment than for the corresponding fixed sample design. If a group sequential trial is stopped early for futility, the patients in the trial do not have to be exposed to potential side effects of the drugs under investigation. Moreover, the resources that would have been required to complete the trial can instead be used to study another treatment in an area of unmet medical need.

- The major administrative benefits can be achieved if an interim analysis is performed early in the trial. It is then possible to check that various aspects of the experiment are running as planned. Problems that are encountered at an early interim analysis might then be dealt with early on, without compromising the entire trial.

- Group sequential methods have the potential to reduce the expected sample size of a clinical trial to approximately $50 - 70\%$ of the corresponding fixed sample design (Jennison and Turnbull, 2000). The possible savings might be compromised if there are requirements to collect a minimum amount of safety data, in which case early stopping for benefit would not make sense until the minimum amount of safety had been collected. The time gained by the early stopping may be important in getting the drug approved sooner, giving a longer time on the market with a valid patent. Morever, patients will be able to benefit from a new effective treatment sooner than if no early stopping would have been possible.

The ethical reasons for interim monitoring through group sequential methods are arguably the most important and persuasive. Most major trials now have a data monitoring committee (DMC), whose primary responsibility is to protect the safety of the patients. ICH guideline E9 Statistical Principles for Clinical Trials encourages the use of interim monitoring through group sequential methods (ICH, 1998). When conducting a group sequential design, there are many aspects to consider that may differ from fixed sample trials. Additional guidelines, that are more specific to interim monitoring, have therefore also been published by regulatory authorities (FDA, 2006). The book by Ellenberg et al. (2003) focuses on practical aspects of interim monitoring, in particular the role of DMCs. Other books more focused on the statistical aspects are written by Whitehead (1997), Jennison and Turnbull (2000) and Proschan et al. (2006). The latter book also covers some of the adaptive methods that will be described next.

1.3 Adaptive designs

1.3.1 Motivation

Over the last decade, the development of new drugs has become increasingly difficult. Fewer new medicines are approved and there are increasing demands on proving that drugs are both safe and efficacious. Approval of treatments for type 2 diabetes are an example of the latter, as the FDA (2008) now requires drug companies to establish both glucose control and that there is not an increased risk of cardiovascular events, for drugs to be approved. In its Critical Path White Paper, the FDA (2004) expresses concerns about the ability of trial sponsors to deal with the increasing problems of getting new drugs approved. From a pharmaceutical industry perspective, older drugs with expiring patents need to be replaced by more effective medicines. More importantly, there are many patients who are waiting for new medicines to treat their conditions. These issues highlight the need for novel ideas for how to develop new and effective medicines. Adaptive designs is one area that has received a lot of attention, both from trial sponsors and regulatory authorities. From the point of view of the trial sponsor, the hope is that this new class of designs will improve the drug development process, for example with issues such as finding the right dose and identifying treatments that should be discontinued. For regulatory authorities, it is helpful if adaptive designs can improve the quality of the clinical trials that they review. Both the EMEA (2007) and the FDA (2010) have issued guidelines about how to apply adaptive methods in clinical trials. Ultimately, the goal of both sponsors and regulators is that a larger number of effective medicines are approved, in particular in areas where there is an unmet medical need.

1.3.2 The breakthrough of adaptive designs

The landmark paper by Bauer and Köhne (1994) opened up new possibilities and laid the foundation for much of the work that has later been done in the area. Bauer and Köhne (1994) consider two sub-samples of patients recruited before and after an interim analysis of a clinical trial. Let the two stages of the trial have the same null hypotheses H_0. The two sub-samples can be used to calculate one-sided p-values p_1 and p_2, associated with H_0. By using Fisher's product criterion, H_0 is rejected if

$$p_1 p_2 \leq C_\alpha = \exp\left[-\frac{1}{2}\chi^2_4(1-\alpha)\right] \tag{1.3}$$

where $\chi^2_4(1-\alpha)$ is the $100 \times (1-\alpha)\%$ quantile of the chi-squared distribution with four degrees of freedom (Bauer and Köhne, 1994). Provided that the combination rule defined in (1.3) has been specified in the protocol, many types of design modifications can be performed between the two stages, without affecting the type I error rate. Unlike in classical group sequential designs, the design modifications may be based upon the p-value p_1 derived from the first sub-sample. Care is however needed when the adaptation is based on data from patients whose primary endpoint, such as overall survival, is only available after the adaptation. Bauer and Posch (2004) point out that in the survival setting the type I error can be inflated if the adaptation is based on biomarkers or covariates that are correlated with the survival endpoint.

It is also possible to define different hypotheses $H_{0,1}$ and $H_{0,2}$ in the two stages, in which case the test in (1.3) is for the intersection hypothesis $H_{0,1} \cap H_{0,2}$. The method of Bauer and Köhne can be generalised to more than two stages, and the recursive combination tests proposed by Brannath et al. (2002) enable derivation of overall p-values and confidence intervals.

A wide range of publications about adaptive designs soon followed, see for example Proschan and Hunsberger (1995); Fisher (1998); Cui et al. (1999); Lehmacher and Wassmer (1999); Denne (2001); Müller and Schäfer (2001). These methods all ensure control of the type I error rate. Adaptations that have been proposed include modification of sample size, identification of sub-populations that might benefit from the drug, switching the primary objective and selecting a sub-set of doses that will be studied for the remainder of the trial. It goes beyond the scope of this thesis to describe all these methods, but we shall now briefly discuss some that are of particular relevance for this thesis.

1.3.3 The methods of Lehmacher and Wassmer and Cui et al.

Lehmacher and Wassmer (1999) ensure protection of the type I error by using a different combination rule from (1.3). The method can be illustrated by considering a group sequential trial with normal response and K groups of observations collected in K

different stages. Let $\tilde{Z}_k$ denote the standardised statistic based on data from stage k of the trial and let

$$Z_M = \frac{\sum_{k=1}^{M} w_k \tilde{Z}_k}{(\sum_{k=1}^{M} w_k^2)^{1/2}}, \quad M = 1, \ldots, K, \tag{1.4}$$

where $w_1, \ldots, w_K$ are weights fixed prior to the start of the trial. Decisions about whether to stop or continue at analysis k is then based on the test statistic Z_k in (1.4).

Marginally, each $\tilde{Z}_k$ follows a $N(\theta\sqrt{\mathcal{I}_k - \mathcal{I}_{k-1}}, 1)$ distribution, where $\mathcal{I}_k$ is Fisher's information for θ at analysis k and $\mathcal{I}_0 = 0$. In Section 3.3.1 we describe the joint distribution of the sequence of statistics $Z_1, \ldots, Z_K$ in (1.4) in further detail. It turns out that under the null hypothesis of no treatment effect, the only difference between this joint distribution and the standard joint canonical distribution in (1.1) is that the correlation structure is now decided by the pre-specified weights in (1.4), rather than the observed information sequence. Hence, critical values for rejecting the null hypothesis that control the type I error rate can be derived in the same way as for non-adaptive group sequential designs with fixed information sequences, using the weights that are fixed at the outset. This method will be evaluated in Chapter 3, as an alternative to error spending designs.

Cui et al. (1999) consider a classical group sequential design and the issue of how to proceed when the effect size turns out to be smaller than what was assumed in setting power at the design stage. They propose a method that allows for sample size modifications based on the observed treatment effect, while protecting the overall type I error rate. The method essentially turns out to be equivalent to that of Lehmacher and Wassmer (1999). Wang et al. (2001) propose an adaptive method for simultaneous testing of superiority and non-inferiority, where control of type I error is ensured in a way similar to that in the method of Cui et al. (1999). The latter is an example of a method for adapting the sample size to obtain adequate power for a new primary objective. The method of Wang et al. (2001) will be evaluated in detail in Chapter 2, where we present our own method for simultaneous testing of superiority and non-inferiority.

1.3.4 Other adaptive designs

Bretz et al. (2006) give a thorough discussion of so-called seamless phase II / III trials. This topic has received a lot of attention, as it has the potential to improve efficiency in different ways. By including data from the phase II subjects in the final analysis, a more precise estimate of the treatment effect on the phase III endpoint can be achieved. In addition, a jointly written protocol for phase II and phase III has the potential to save time that is otherwise spent preparing and getting approval for the phase III protocol. Seamless phase II / III designs can be thought of as a special case of the more general idea of incorporating data from different phases in a single trial. These ideas seek to

relax the assumption that drug development has to be rigidly divided into the phases described in Section 1.1. There is also a connection to the problem of jointly planning phase II and phase III, considered in Chapter 5 of this thesis. Koch (2006) and the EMEA (2007) express some caution about the use of seamless phase II/III designs, but recognise that such designs could be of value if appropriately planned.

We shall refer to classical group sequential designs as non-adaptive group sequential designs, although some authors like Dragalin (2006) consider group sequential testing as a form of adaptation. They make it possible to stop a trial early, according to the stopping rule that is defined by the group sequential boundary. By using the error spending method of Lan and DeMets (1983), the type I error can be controlled when future group sizes are unpredictable. Future group sizes in an error spending test should however not be chosen based on the observed treatment effect, as this can lead to inflation of the type I error rate (Proschan et al., 1992). The situation is different for adaptive methods, where this and other design changes can be accommodated while maintaining type I error control. As it is possible to add group sequential boundaries to an adaptive design with interim analyses, adaptive designs can be viewed as a generalisation of non-adaptive group sequential designs. Several papers have discussed the relative efficiency of non-adaptive and adaptive group sequential designs, with a focus on methods with sample size modification based on the observed treatment effect. Jennison and Turnbull (2006a) showed that an optimal non-adaptive group sequential design can come within a few percentage points of the efficiency of an optimal adaptive group sequential design. Adaptive designs can on the other hand be substantially inferior if a sub-optimal rule for sample size modification is applied (Jennison and Turnbull, 2003). Burman and Sonesson (2006) criticised adaptive methods from another perspective, arguing that some adaptive designs may lack credibility and be difficult to interpret. Burman and Lisovskaja (2010) proposed to address this through the so-called dual test, where it is required that both the adaptive test and a naïve test, ignoring the adaptations, are statistically significant.

Müller and Schäfer (2001, 2004) proposed a new adaptive method referred to as the conditional rejection probability (CRP) principle. It is a very general method that gives almost complete flexibility for how to re-design a clinical trial. The method is very appealing as it tries to combine the benefits of classical group sequential designs with novel adaptive approaches. First, a group sequential design is set up to have all the benefits of sequential monitoring. Thereafter, an adaptive re-design is applied only when considered necessary, based on all available internal and external information. The CRP principle is discussed in detail in Chapter 3, where we show that the method has some limitations when applied in an error spending design with unpredictable group sizes.

1.4 Decision Analysis

In the previous sections we have referred to various group sequential and adaptive methods. Some of these methods have optimal properties in the sense that they minimise the expected sample size, while satisfying suitably defined frequentist error probability constraints. Many of these optimal designs were derived using decision analysis. A key feature of decision analysis is to define a utility function that should be maximised, or equivalently, a loss function that should be minimised. Given data, one would like to find the action or decision rule that is optimal, in the sense that it maximises the expected utility.

Our focus will be on how to apply decision analysis to derive efficient clinical trial designs. It is however important to recognise that decision analysis has many applications beyond clinical trials. The book by Berger (1985) is a standard reference about the subject. It is focused on the underlying theory but also includes a few examples from different areas. Berger points out that there is a strong link between decision analysis and Bayesian statistics. Apart from defining a utility function and a cost of sampling, the formulation of a decision problem typically involves choosing a prior distribution for the unknown parameter. The link to Bayesian ideas is emphasised by Lindley (1997), who argues for choosing the sample size of an experiment based on Bayesian decision analysis. The book by Parmigiani and Inoue (2009) is another useful reference that provides theoretical background as well as applications in areas such as biostatistics and economics.

There are plenty of examples where decision analysis has been used in work related to clinical trials, in different phases of drug development. Stallard (1998) uses cost and utility functions to derive optimal group sequential phase II designs for binary outcomes. Gittins and Pezeshk (2000b) derive the optimal sample size of a phase III trial by modeling how the number of patients using the drug depends on the posterior distribution of the treatment effect after the trial. Burman et al. (2007, Chapter 14) give a good overview about the use of decision analysis in drug development. They emphasize how decision analysis can be applied to a wider range of problems than the design of a single trial, for example project prioritisation. Julious and Swank (2005) provide another interesting example of how decision analysis can be used in a wider context. They describe how decision analysis based on elicitation of expert knowledge can be used to choose between different options for a clinical development plan. We refer to the book by O'Hagan et al. (2006) for a good discussion about how elicitation can be used to determine an appropriate prior distribution.

Decision analysis can also be used to solve the problem of finding an optimal group sequential boundary. The possible actions are then typically to stop and make a decision about the unknown parameter or to continue sampling. The sequential nature of the problem means that the method of dynamic programming can sometimes be applied.

When using dynamic programming to derive an optimal sequential decision rule, one makes use of the fact the optimal decision rule at the final analysis does not depend on the decision rules at previous analysis. When searching for an optimal group sequential boundary, it is consequently possible to find the optimal critical values at the final analysis and work backwards to the first analysis. This works because it is possible to consider an unconstrained decision problem, and thereafter search for the costs that give the desired frequentist error probability constraints. Dynamic programming has been used to derive optimal group sequential boundaries for a variety of problems, see for example Lai (1973); Eales and Jennison (1992, 1995); Barber and Jennison (2002); Hampson (2009). A good introduction to the method is given in the book by Bather (2000).

In Chapter 4, we solve the problem of finding optimal group sequential designs with non-binding futility boundaries. This is an example of a decision problem where dynamic programming cannot be directly applied, as the optimal decision rule at the final analysis depends on the critical values at previous analyses. We nevertheless show that by extending the method of dynamic programming, this more complex problem can be solved.

1.5 Thesis organisation and road map

In this thesis we are concerned with using sequential and adaptive methods to achieve efficient clinical trial designs. In three of the problems we have formulated a utility that we seek to maximise, using decision analysis to derive an efficient design. Apart from deriving efficient group sequential designs, we are interested in the increased benefits that can be achieved through adaptive methods. We also address potential issues with adaptive methods, such as control of type I error and the impact of sub-optimal rules for sample size modification.

Chapter 2 is mainly based on the publication by Öhrn and Jennison (2010). The research for this paper was carried out primarily by Öhrn, but the paper was written jointly by Öhrn and Jennison. The problem was motivated by the adaptive methods described in Section 1.3 and the idea of adaptively switching primary objective. We derive optimal group sequential designs for simultaneous testing of superiority and non-inferiority and compare their efficiency to designs that have been proposed in the literature. To assess the benefits of sample size modification based on the observed treatment effect, efficient adaptive group sequential designs are derived and compared with their non-adaptive counterparts. It is found that both objectives can be addressed within a non-adaptive group sequential framework, while the additional benefits achieved through sample size modification are modest. We also define error spending versions of the non-adaptive group sequential designs that can be used to

cope with unpredictable group sizes and information levels.

In Chapter 3 we focus on the CRP principle, one of the adaptive methods discussed in Section 1.3. The method seeks to combine the benefits of error spending designs with novel adaptive approaches. It is found that when applying the CRP principle, there is an issue related to type I error control that has to be dealt with. There are ways to make sure that the type I error rate is controlled, but some of the generality of the approach is then lost. It is concluded that rather than enjoying all the benefits of error spending design, the CRP principle provides some of these benefits, as well as additional advantages related to adaptivity.

In Chapter 4 we are concerned with one-sided group sequential tests with non-binding futility boundaries. The need for futility boundaries that are non-binding is an issue that has often been overlooked in the literature about group sequential designs. In Chapter 4 we assess the implications of this problem, which is caused by practical aspects of interim monitoring and how regulatory authorities view the lower boundary. A new method for deriving optimal group sequential designs with non-binding futility boundaries is presented. Existing designs with non-binding futility boundaries are reviewed and compared to designs derived with the new method.

In Chapter 5 we move beyond the individual trial and try to assess how the design of a series of trials can be approached. In a very general framework, we discuss how to jointly optimise two phases of a development program. Making use of the knowledge gained in Chapter 4, we also assess how the properties of the trial programme change if a group sequential design with a non-binding futility boundary is applied in phase III. The problem with two studies in Chapter 5 bears similarities with the optimisation of a group sequential design with two groups. The development programme can be stopped after phase II, which can be thought of as stopping for futility in the group sequential design. If the phase III sample size is chosen based on phase II data, the problem becomes similar to the so-called optimal sequentially planned decision procedures proposed by Schmitz (1993). A version of the Schmitz designs adapted to a three-decision problem is studied in Chapter 2.

All of the chapters in this thesis deal with sequential and adaptive methods for the design of clinical trials. We have used computational methods to derive efficient designs for a variety of problems, with a special focus on the benefits that can be achieved through group sequential and adaptive methods. In Chapter 6 this thesis is finished by discussing general conclusions that can be drawn and outlining possible areas for future research.

CHAPTER 2

Optimal group sequential designs for simultaneous testing of superiority and non-inferiority

2.1 Introduction

The primary objective in many clinical trials is to demonstrate superiority of the experimental treatment. With an active control treatment, it may also be of interest to show the experimental treatment is not worse than the control by more than a pre-specified margin. Proving "non-inferiority" is particularly appropriate if the new treatment is safer than the control.

Let θ denote the treatment difference between the new treatment and control, with positive values of θ indicating superiority of the new treatment. Superiority can be established by rejecting the null hypothesis $H_{S,0}$: $\theta \leq 0$ in favour of the alternative $\theta > 0$. Suppose it is agreed that the new treatment may be regarded as non-inferior if $\theta > -\delta_N$, where δ_N is a positive quantity referred to as the non-inferiority margin. We shall conclude that the new treatment is non-inferior if the null hypothesis $H_{N,0}$: $\theta \leq -\delta_N$ is rejected in favour of $\theta > -\delta_N$.

Morikawa and Yoshida (1995) note that tests for superiority and non-inferiority involve nested hypotheses and, hence, overall type I error probability will be controlled if both tests are conducted simultaneously without any adjustment for multiplicity. The same is true if the two hypotheses are tested group sequentially in a closed testing procedure (Wang et al., 2001). However, the sample sizes required for tests of superiority and non-inferiority may be quite different. Suppose the test for non-inferiority is to have type I error probability α at $\theta = -\delta_N$ and power $1 - \beta$ at $\theta = 0$, while the test for superiority has type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta_S$. The value of δ_N is typically set as a fraction of the estimated treatment difference in an earlier comparison of the active control treatment and placebo, and is

liable to be quite small. The value of δ_S may be chosen to reflect expectations of a substantial treatment effect and when this is significantly larger than δ_N, the sample size needed for the test of non-inferiority will be considerably higher than that required to test for superiority.

The need for different sample sizes to test the two hypotheses has led to quite complex proposals for group sequential designs testing both superiority and non-inferiority. Wang et al. (2001) describe an adaptive group sequential procedure in which the sample size is initially set for a test of superiority but, if interest shifts to showing non-inferiority, group sizes are increased. When this data-dependent change occurs, the type I error rate is preserved by down-weighting later groups of observations in the manner of Cui et al. (1999).

In the two-stage procedures of Shih et al. (2004) and Koyama et al. (2005), first stage data are used to decide whether to continue and, if so, to select superiority or non-inferiority as the primary objective. The second stage sample size is chosen to give power for the chosen objective: Shih et al. (2004) set sample size as a function of first stage data to achieve a given conditional power, while Koyama et al. (2005) use a sample size function attaining a specified unconditional power.

Lai et al. (2006) describe non-adaptive group sequential designs with fixed group sizes and three possible decisions on termination: superiority, non-inferiority (but not superiority) and inferiority. Reaching the third decision, inferiority, is sometimes referred to as stopping for futility since there is little prospect of reaching either positive decision. When δ_S is greater than δ_N, the study can terminate at an early stage with a decision of superiority or inferiority then, later on, the options switch to non-inferiority and inferiority.

We shall present general classes of group sequential procedures which build on existing proposals. We first discuss designs with fixed group sizes, extending the options considered by Lai et al. (2006) by allowing a choice of all three terminal decisions at each analysis. In our formulation of the testing problem in Section 2.2, values for δ_S and δ_N are stipulated along with a type I and type II error probability for each hypothesis test. For a given sequence of group sizes, we derive designs with the lowest possible expected sample sizes averaged over a range of values of the treatment effect, θ, while meeting the error probability constraints. Although group sizes are fixed, these procedures still exhibit a form of adaptation: when δ_S is significantly greater than δ_N, the upper continuation region for testing between superiority and non-inferiority comes to an end first, while the lower region continues to allow differentiation between non-inferiority and inferiority.

In Section 2.3 we generalise these designs to let group sizes depend on previously observed data. The resulting class includes the adaptive group sequential designs of Wang et al. (2001) and the adaptive two-stage procedures of Shih et al. (2004) and

Koyama et al. (2005). In the two-decision problem of a one-sided test for superiority, Jennison and Turnbull (2006a) found that adaptive choice of group sizes provided only a slight efficiency gain over non-adaptive designs. In our three-decision problem, when different fixed sample sizes are appropriate to the two separate hypothesis tests, it seems plausible that there could be more substantial gains from using interim data both to choose the null hypothesis on which to focus and to adjust sample size accordingly. We assess previously proposed designs and new, optimised two-stage procedures to investigate the reduction in expected sample size that can be achieved by such adaptation. Our conclusion from the examples we have studied is that little is gained by choosing the second group size based on the observed treatment effect.

2.2 Optimal non-adaptive designs

2.2.1 Framework

Suppose that the observations X_{Aj} and X_{Bj}, $j = 1, 2, \ldots$, on treatments A and B, respectively, are independent and normally distributed with $X_{Aj} \sim N(\mu_A, \sigma^2)$ and $X_{Bj} \sim N(\mu_B, \sigma^2)$. We assume for now that σ^2 is known but we shall explain in Section 2.2.5 how unknown variance can be handled. The parameter of interest is the treatment effect $\theta = \mu_B - \mu_A$. We wish to test simultaneously $H_{N,0}$: $\theta \leq -\delta_N$ against $\theta > -\delta_N$ and $H_{S,0}$: $\theta \leq 0$ against $\theta > 0$, where the non-inferiority margin δ_N is established prior to the start of the trial.

Gould (1997) and Koyama et al. (2005) recognise this is a three-decision problem with outcomes: superiority, non-inferiority (only) and inferiority. Error rate requirements, including power for the two hypothesis tests at $\theta = 0$ and $\theta = \delta_S$, can be expressed through a pair of power curves. The curves displayed in Figure 2-1 show the probabilities of concluding "Non-inferiority or Superiority" or "Superiority" as functions of θ. Formally, we specify type I and type II error rates α_N and β_N for testing $H_{N,0}$ and error rates α_S and β_S for testing $H_{S,0}$ as:

$$P_{\theta=-\delta_N}(\text{Declare "Non-inferiority" or "Superiority"}) = \alpha_N, \tag{2.1}$$

$$P_{\theta=0}(\text{Declare "Superiority"}) = \alpha_S, \tag{2.2}$$

$$P_{\theta=0}(\text{Conclude "Inferiority"}) = \beta_N, \tag{2.3}$$

$$P_{\theta=\delta_S}(\text{Conclude "Inferiority" or "Non-inferiority"}) = \beta_S. \tag{2.4}$$

In Section 2.6.1 we prove that these conditions imply control of type I error for $H_{N,0}$ and $H_{S,0}$ over all values $\theta \leq -\delta_N$ and $\theta \leq 0$, respectively, and of type II error over $\theta \geq 0$ and $\theta \geq \delta_S$. This result holds for all non-adaptive designs satisfying certain minimal criteria, and it also applies to the adaptive designs we shall introduce in Section 2.3.

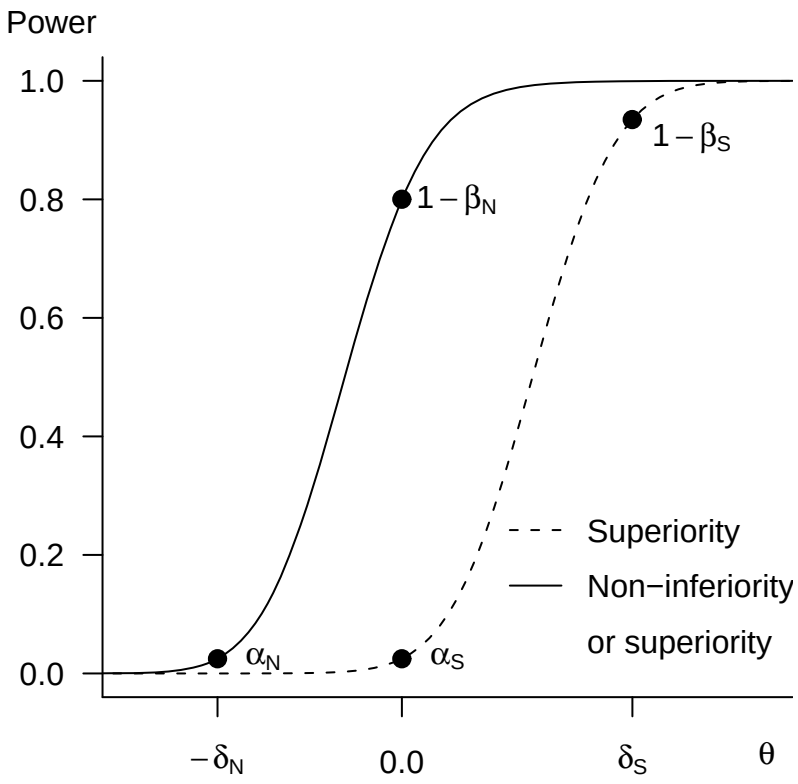

Figure 2-1: Power curves for "non-inferiority or superiority" and superiority.

For many designs, fixing two points on the power curve results in the whole curve being indistinguishable from that of a fixed sample test. Hence, we do not consider power under other parameter values when comparing designs. The exceptions are some adaptive designs for which power approaches unity rather slowly: see, for example, the power curves in Figure 2-11.

If the tests of the two null hypotheses were carried out in separate fixed sample trials, the number of observations required per treatment would be

$$n_{Nf} = 2\{\Phi^{-1}(1-\alpha_N) + \Phi^{-1}(1-\beta_N)\}^2\sigma^2/\delta_N^2$$

for testing $H_{N,0}$ and

$$n_{Sf} = 2\{\Phi^{-1}(1-\alpha_S) + \Phi^{-1}(1-\beta_S)\}^2\sigma^2/\delta_S^2$$

for testing $H_{S,0}$, where Φ is the cumulative distribution function of a standard normal variate.

We shall consider group sequential procedures with a maximum of K analyses, denoting the cumulative sample size per treatment at analysis k by n_k and the maximum sample size per treatment by $n_{max} = n_K$. Let Z_k be the standardised

test statistic for testing $\theta = 0$ at analysis k. Allowing early stopping for all possible decisions at each analysis leads to a rule at analysis k of the form:

if $Z_k \leq a_k$,	stop and conclude inferiority,
if $a_k < Z_k < b_k$,	continue sampling,
if $b_k \leq Z_k \leq c_k$,	stop and declare non-inferiority,
if $c_k < Z_k < d_k$,	continue sampling,
if $Z_k \geq d_k$,	stop and declare superiority.

Here, $a_k \leq b_k \leq c_k \leq d_k$ and termination by analysis K is ensured by setting $a_K = b_K$ and $c_K = d_K$. At the final analysis, superiority is declared if $Z_K \geq d_K$, while non-inferiority is declared if $b_K \leq Z_K < d_K$. When $n_{Sf} < n_{Nf}$, we may have $c_k = d_k$ in later stages so the upper continuation region is not present. In such cases, we denote the first value of k at which $c_k = d_k$ by K_S and the planned group size at this analysis by $n_{max,S}$. Although the lower boundary a_k, is present throughout, it can be helpful to think of the design as focusing on the test for superiority up to analysis K_S and concentrating on the choice between non-inferiority and inferiority thereafter. In order to meet the error probability constraints, it will be necessary for $n_K = n_{max}$ to be greater than n_{Nf} and $n_{K_S} = n_{max,S}$ to be greater than n_{Sf}. We shall refer to the ratios

$$r_S = n_{max,S}/n_{Sf} \quad \text{and} \quad r_N = n_{max}/n_{Nf}$$

as "inflation factors" and use these to indicate how much the maximum sample size, for the first phase or the whole design, has been increased beyond the minimum requirement.

A typical rule is illustrated in Figure 2-2. In this example, $b_k = c_k$ for $k = 1$ and 2 so early stopping for non-inferiority is not possible at the first two analyses and the two sections of continuation region merge into one. Since $c_k = d_k$ for $k = 4$ to 8, there is no upper continuation region at these analyses. However, we still allow the possibility to stop with a conclusion of superiority if the last group of observations results in a sufficiently high value of Z_5, Z_6, Z_7 or Z_8. The boundaries in Figure 2-2 are those of an optimal design which we shall describe in Section 2.2.2. They are broadly similar to the two-sided tests with an inner wedge described by (Jennison and Turnbull, 2000, Chapter 6), however, they lack the symmetry of those designs around $\theta = 0$ and the roles of two of the error probabilities, α_N and β_N, are reversed.

The boundary points $a_1, b_1, c_1, d_1, \ldots, a_K, b_K, c_K, d_K$ must be chosen to satisfy the error constraints (2.1) to (2.4). A fixed sample size trial can only meet all four constraints simultaneously if $n_{Nf} = n_{Sf}$. In contrast, the additional degrees of

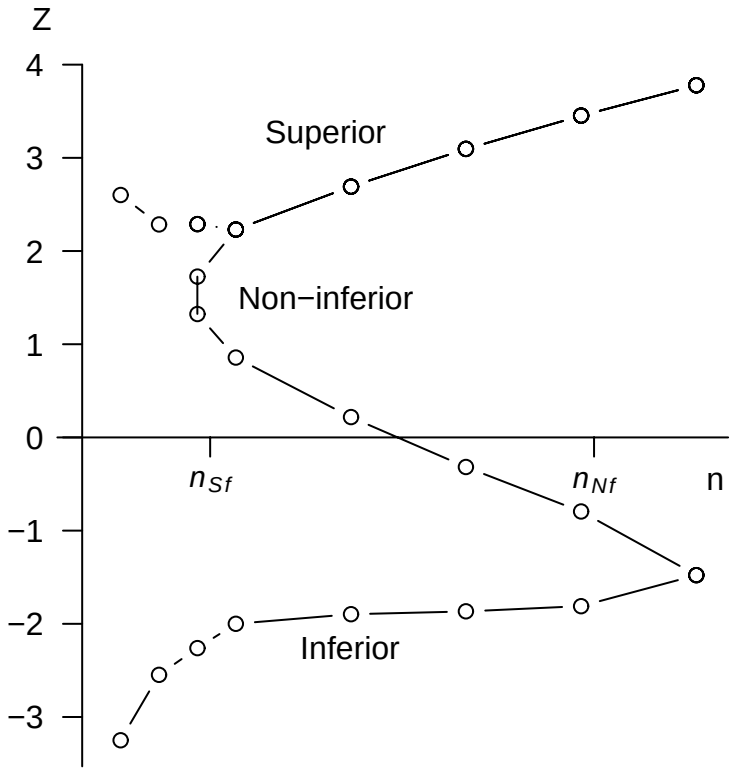

Figure 2-2: Stopping boundaries for inferiority, non-inferiority and superiority.

freedom of a group sequential design allow suitable boundaries to be found as long as $n_{max} > \max(n_{Nf}, n_{Sf})$. Moreover, we can exploit the remaining degrees of freedom to find a trial design with low expected sample size under specified values of θ. We have developed methods to find group sequential designs that minimise criteria of the form $\sum_{i=1}^{m} w_i E_{\theta_i}(N)$, where N denotes the sample size per treatment on termination. We have studied designs for a variety of optimality criteria, but in this thesis we shall focus on

$$F^\star = \{E_{-\delta_N}(N) + E_{-\delta_N/2}(N) + E_0(N) + E_{\delta_S/2}(N) + E_{\delta_S}(N)\}/5 \tag{2.5}$$

which combines performance across the range of effect sizes of interest. We shall illustrate these procedures with an example for particular design parameters in Section 2.2.3 and make an efficiency comparison with the design of Lai et al. (2006) in Section 2.2.4.

2.2.2 Derivation of optimal designs

Our methods enable us to find an optimal design with a specified number of analyses K and cumulative sample sizes $n_1, \ldots, n_K$. Comparing these optimal designs for different

sequences $n_1, \ldots, n_K$ can inform the choice of suitable group sizes. Increasing K will decrease expected sample size but at the cost of more interim analyses, so comparing optimal designs for different values of K helps assess their costs and benefits.

Our derivation of optimal designs extends the methods of Eales and Jennison (1992), Eales and Jennison (1995) and Barber and Jennison (2002) to the asymmetric three-decision problem. Given K and $n_1, \ldots, n_K$, we seek the stopping boundary minimising $\sum_{i=1}^{m} w_i E_{\theta_i}(N)$ subject to error probability requirements (2.1) to (2.4). We follow a Lagrangian approach and introduce the unconstrained problem of minimising

$$\sum_{i=1}^{m} w_i E_{\theta_i}(N) + \lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3 + \lambda_4 P_4, \tag{2.6}$$

where λ_1 to λ_4 are positive and P_1 to P_4 denote the left hand sides of equations (2.1) to (2.4). The design minimising (2.6) must have the minimum value of $\sum_{i=1}^{m} w_i E_{\theta_i}(N)$ among all designs with the same P_1 to P_4. Hence, choosing Lagrange multipliers λ_1 to λ_4 so that the solution has $P_1 = \alpha_N$, $P_2 = \alpha_S$, $P_3 = \beta_N$ and $P_4 = \beta_S$ solves the original constrained problem.

For given λ_1 to λ_4, the method of dynamic programming can be used to minimise (2.6) quickly and accurately. This minimisation problem also has an interpretation as a Bayes sequential decision problem with a certain combination of prior on θ, costs for incorrect decisions, and sampling costs under the θ_is appearing in $\sum_{i=1}^{m} w_i E_{\theta_i}(N)$. This Bayesian interpretation provides insight into the dynamic programming solution where it is seen that decisions at each stage are based on expected future costs under the current posterior distribution for θ. Further details of the derivation of optimal designs are provided in Section 2.6.2.

2.2.3 Numerical example

Suppose a non-inferiority margin of $\delta_N = 0.1$ has been established, power for the test of superiority is set at $\delta_S = 0.2$, and error probabilities are $\alpha_N = 0.025$, $\alpha_S = 0.025$, $\beta_N = 0.1$, and $\beta_S = 0.1$. If the response variance is $\sigma^2 = 0.5$, the fixed sample size per treatment for the test of superiority alone is $n_{Sf} = 263$, while the test for non-inferiority needs $n_{Nf} = 1051$.

We first consider a design with $K = 8$ analyses. The maximum sample size, n_{max}, must be at least a little greater than the larger of n_{Sf} and n_{Nf}, so we choose $n_{max} = 1.2\, n_{Nf} = 1261$. We set $n_4 = n_{max,S} = 1.2\, n_{Sf} = 316$ so a conclusion about the superiority objective can be reached in the first four analyses, leaving analyses 5 to 8 to concentrate on testing between non-inferiority and inferiority. Taking equal group sizes either side of analysis 4, we have $n_k = (k/4)n_4$ for $k = 1, \ldots, 3$, and $n_k = n_4 + ((k-4)/4)(n_{max} - n_4)$ for $k = 5, \ldots, 8$.

Optimising the design for $F^\star$ yields the boundary values a_1, b_1, c_1, $d_1, \ldots, a_8$, b_8,

c_8, d_8 shown earlier in Figure 2-2. The absence of an inner wedge at the first two analyses indicates it is not possible to stop early for non-inferiority in this optimal design. The form of the upper part of the stopping boundary at later analyses merits some comment. Since optimisation has produced $c_k = d_k$ for $k = 4$ to 8, there is no upper continuation region after analysis 4; this is in line with our intent to deal with the issue of superiority by this analysis. The presence of values for d_5 to d_8 shows it is still possible to decide in favour of superiority at a later analysis if the last group of observations produces a large increase in the Z-statistic and the study ends with $Z_k > d_k$. In fact, such a sequence of Z_ks is unlikely under any value of θ and the dramatic change in observed treatment effect in the final group needed to achieve this might well raise questions about heterogeneity of the treatment effect over time. Let K_S denote the index of the first analysis at which $c_k = d_k$ in an optimal design. We have found that setting $c_k = d_k = \infty$ for $k > K_S$, so that only the test between non-inferiority and inferiority is pursued at analyses $k = K_S + 1, \ldots, K$, has a negligible impact on error probabilities and, hence, on efficiency. One may, therefore, choose to remove the option of a decision in favour of superiority after the analysis at which values of c_k and d_k first converge. This will be the case in our definition of error spending designs in Section 2.2.5. However, unless otherwise stated we shall retain the option of stopping for superiority, and finite values for the d_k, in the optimal designs we report.

Hence, we can have a trial that stops for non-inferiority (but not superiority) at analysis k, if $b_k \leq Z_k \leq c_k$, while for another trial we can have $Z_k < b_k$ but still claim superiority later on, if the superiority boundary is crossed at a later analysis. As discussed in Section 2.6.1, this property makes the proof of monotonicity of type I and type II error probabilities more involved. In principle, it would have been possible to follow the approach of Lai et al. (2006) and not allow a superiority decision at later analysis. We believe however that it is sensible to declare superiority if the estimated treatment effect at the end of the trial is overwhelmingly in favour of such a decision, even though results were less impressive at an earlier interim analysis. In practice, it is likely to be important to perform sensitivity analyses to understand the reasons for the heterogeneity of the treatment effect over time.

The optimal design's expected sample size is shown as a function of θ by the solid line in Figure 2-3. The two horizontal lines at n_{Nf} and n_{Sf} aid comparison with the sample sizes of the individual, fixed sample tests for non-inferiority and superiority. The sequential design is clearly effective in reducing expected sample size below n_{Nf}. Since the fixed sample size, n_{Sf}, in the individual test for superiority is insufficient to provide the desired power for the test of non-inferiority, it is to be expected that the sequential design has expected sample size greater than n_{Sf} at low values of θ. However, at high values of θ, where the main task is to distinguish between superiority and non-inferiority, $E_\theta(N)$ does fall below n_{Sf}. Additional curves in Figure 2-3 show the

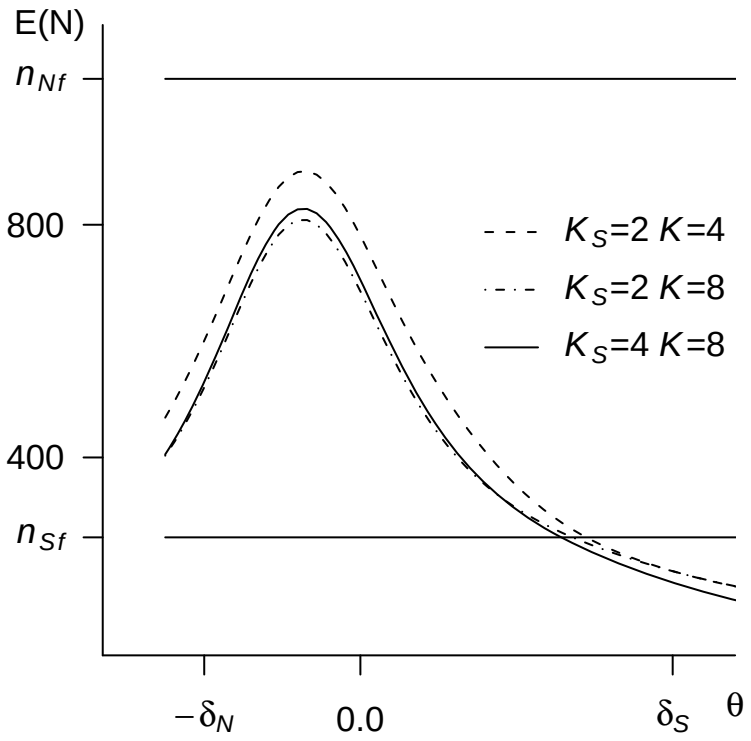

Figure 2-3: Expected sample size functions for optimal designs with 4 and 8 analyses.

expected sample size function for two more optimal designs, one with $K = 4$ analyses and one with $K = 8$ analyses. For both these designs, $K_S = 2$, $n_2 = 1.1\, n_{Sf} = 289$, and the remaining $K - 2$ analyses are equally spaced up to $n_K = 1261$. With a total of 8 analyses, reducing K_S from 4 to 2, so that more analyses are devoted to the inferiority/non-inferiority comparison, reduces $F^\star$ by about 1%. However, the design with $K_S = 4$ performs better for large values of θ. The efficiency gained by increasing the total number of analyses from 4 to 8 is close to 10%, a larger improvement than is typically seen in one-sided group-sequential tests of $\theta \leq 0$ against $\theta > 0$. This can be attributed to the fact that some analyses are well placed for one testing objective but poorly placed for the other, so the "effective" number of analyses for testing each individual hypothesis is less than K. We conclude that when the ratio n_{Nf}/n_{Sf} is as high as 4, designs with only a small number of groups may not achieve the full benefits of sequential monitoring.

2.2.4 Comparison with the method of Lai et al.

Lai et al. (2006) propose a group sequential procedure with K analyses that switches objective at a certain analysis. For $n_{Nf} > n_{Sf}$, the procedure allows early stopping

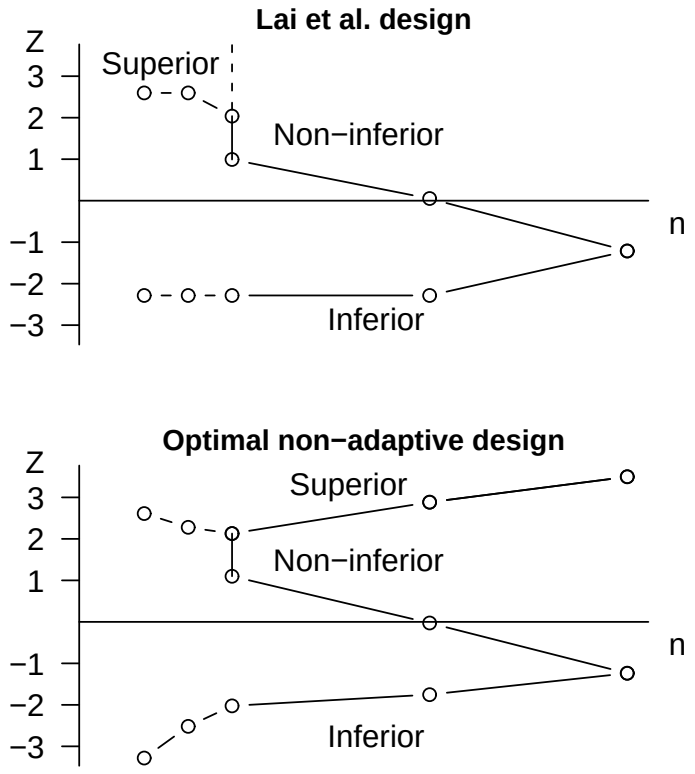

Figure 2-4: Critical values for a Lai et al. 5-group design and an optimal 5-group design.

for superiority at analyses $1, \ldots, K_S$ and for non-inferiority at analyses $K_S, \ldots, K$; the test can stop for the negative decision of inferiority at any point. The authors present their method in terms of generalized likelihood ratio statistics but we shall define critical values on the Z scale. Figure 2-4 displays a 5-group Lai et al. procedure with $K_S = 3$. Note that decisions of non-inferiority (solid line) and superiority (dashed line) are both possible at analysis 3. The authors define a parameter ϵ governing the amount of early stopping and recommend using $\epsilon = 1/3$. The outer boundaries have constant critical values on the Z scale so, in our notation, $d_1 = \ldots = d_{K_S-1} = \tilde{d}$ and $a_1 = \ldots = a_{K-1} = \tilde{a}$. For the non-inferiority boundary, $b_k = \tilde{b} - \delta_N \sqrt{\{n_k/(2\,\sigma^2)\}}$ for $k = K_S, \ldots, K-1$. The values of $\tilde{a}$, a_K, $\tilde{b}$, $\tilde{d}$ and d_{K_S} are chosen so that: the probability under $\theta = 0$ of stopping to declare superiority by analysis $K_S - 1$ is $\epsilon\,\alpha_S$ and by analysis K_S is α_S; the probability under $\theta = -\delta_N$ of stopping to declare non-inferiority or superiority by analysis $K-1$ is $\epsilon\,\alpha_N$ and by analysis K is α_N; the probability under $\theta = 0$ of stopping to conclude inferiority by analysis $K-1$ is $\epsilon\,\beta_N$.

This construction does not produce specific type II error probabilities, instead these are determined by n_K, n_{K_S} and ϵ. There is a demarcation at analysis K_S, with stopping

for superiority only possible at analyses 1 to K_S, and stopping for non-inferiority only at analyses K_S to K. The framework of Section 2.2 imposes no such constraints and we allow an inner wedge for non-inferiority before K_S and a continuing superiority/non-inferiority boundary thereafter.

We have applied the method of Lai et al. to the numerical example of Section 2.2.3, in which $\delta_N = 0.1$, $\delta_S = 0.2$, $\sigma^2 = 0.5$ and $\alpha_N = \alpha_S = 0.025$. We set $K = 5$, $K_S = 3$ and $\epsilon = 1/3$, with $n_{max,S} = 263$ and $n_{max} = 1051$, the values that would give 90% power if the two testing objectives were addressed in fixed sample trials. This implies $n_1 = 88$, $n_2 = 175$, $n_3 = n_{max,S} = 263$, $n_4 = 657$, and $n_5 = n_{max} = 1051$. It is the resulting boundaries that are shown in the upper panel of Figure 2-4. This design has type II error probabilities $\beta_N = 0.125$ and $\beta_S = 0.11$. Using these numbers to define fixed sample sizes n_{Sf} and n_{Nf}, we find the inflation factors for the Lai et al. design are $r_S = 1.035$ and $r_N = 1.086$.

We compared the Lai et al. (2006) design with a 5-group sequential design with the same group sizes and attained error probabilities, optimised for $F^\star$. This optimised design with $K_S = 3$, $r_S = 1.035$ and $r_N = 1.086$ is depicted in the lower panel of Figure 2-4 and its expected sample size function is shown in Figure 2-5: the value of $F^\star$ is about 4% lower than that of the Lai et al. design. However, the inflation factor $r_S = 1.035$ is rather low and there is no obvious need to restrict $n_{max,S}$, given that higher sample sizes are allowed if the study continues to later analyses. Keeping $K_S = 3$ and increasing r_S from 1.035 to 1.2 while maintaining the same overall maximum sample size gives cumulative group sizes $n_1 = 102$, $n_2 = 204$, $n_3 = 306$, $n_4 = 678$, and $n_{max} = 1051$. The boundary optimising $F^\star$ for these group sizes has an inner wedge at the second interim analysis. It is evident from Figure 2-5 that this increase in r_S reduces the expected sample size function a little. In the example of Section 2.2.3, we found some advantage in scheduling fewer analyses for the superiority objective, leaving more to test between non-inferiority and inferiority. Here, we have considered $K_S = 2$ and $r_S = 1.2$, so $n_1 = 153$, $n_2 = 306$, $n_3 = 554$, $n_4 = 803$ and $n_5 = 1051$. The lowest curve in Figure 2-5 is for the test minimising $F^\star$ with these group sizes and we see this design has the smallest expected sample size at all but the highest effect sizes.

Overall, we recognise that Lai et al's proposal gives designs with quite good efficiency, but our wider class allows a design to be tailored to particular objectives and the "inner wedge", not considered by Lai et al, can be instrumental in reducing expected sample size.

2.2.5 Error spending designs

To be of real practical value, a group sequential method should be able to deal with variation in group sizes about their planned values. Error spending designs offer this flexibility and, we shall show, can do so with high efficiency in terms of

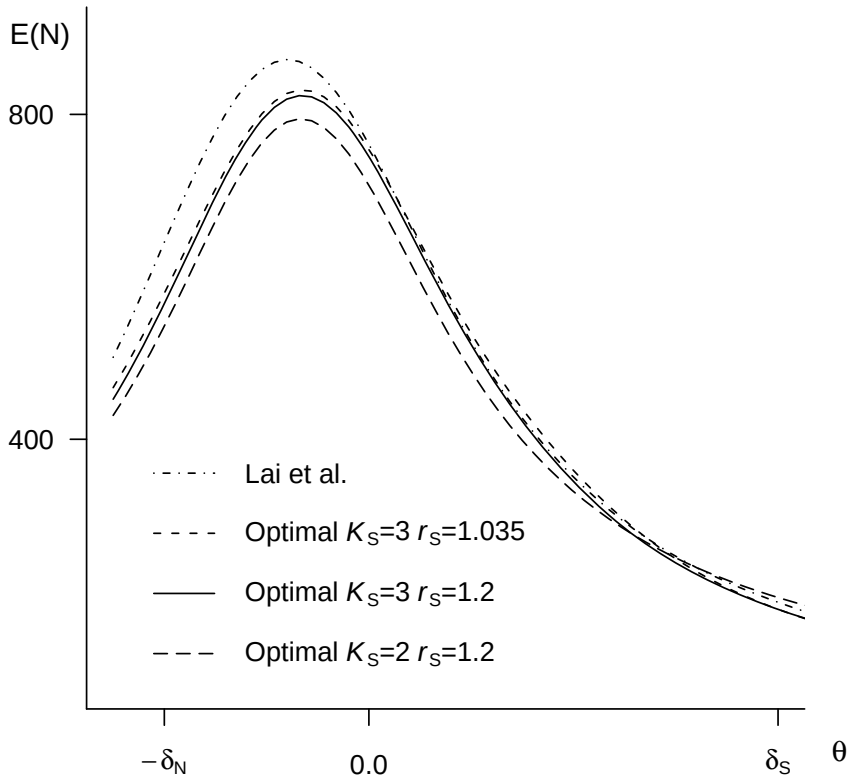

Figure 2-5: Expected sample size functions for the Lai et al. design and three optimal 5-group designs.

expected sample size. In introducing these designs we broaden consideration to general response distributions, still with the parameter θ representing the treatment effect under investigation. Jennison and Turnbull (1997) show that for normal linear models, and asymptotically for general parametric models, the sequence of estimates $\hat{\theta}_1, \ldots, \hat{\theta}_K$ based on accumulating data at K analyses is multivariate normal with

$$\hat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1}), \quad k = 1, \ldots, K,$$

and

$$\mathrm{Cov}(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) = \mathrm{Var}(\hat{\theta}_{k_2}) = \mathcal{I}_{k_2}^{-1} \quad \text{for } k_1 < k_2,$$

where $\mathcal{I}_k$ represents the Fisher information for θ at analysis k.

In error spending designs, the cumulative type I and type II error probabilities are specified as functions of the observed information at each analysis; boundaries for standardised statistics $Z_k = \hat{\theta}_k \sqrt{\mathcal{I}_k}$ are found that satisfy these conditions using the distribution theory stated above. This approach relaxes the requirement of achieving pre-planned information levels at each analysis implicit in the tests of Section 2.2. Lan

and DeMets (1983) introduced error spending to handle unpredictable group sizes in two-sided tests of a null hypothesis. We now extend this approach to our three-decision problem with its four type I and type II error probabilities.

The information levels required by individual tests of superiority and non-inferiority are

$$\mathcal{I}_{Sf} = \{\Phi^{-1}(1-\alpha_S) + \Phi^{-1}(1-\beta_S)\}^2/\delta_S^2$$

and

$$\mathcal{I}_{Nf} = \{\Phi^{-1}(1-\alpha_N) + \Phi^{-1}(1-\beta_N)\}^2/\delta_N^2.$$

Multiplying these expressions by inflation factors r_S and r_N gives target information levels for an error spending design. Assuming $\mathcal{I}_{Nf} > \mathcal{I}_{Sf}$, the overall maximum information level which an error spending design may require is $\mathcal{I}_{max} = r_N\,\mathcal{I}_{Nf}$. We also specify a target information level by which testing for superiority should terminate, $\mathcal{I}_{max,S} = r_S\,\mathcal{I}_{Sf}$.

Type I and II error probabilities α_S and β_S for the test of superiority are spent according to functions $f_S(\mathcal{I})$ and $g_S(\mathcal{I})$ as $\mathcal{I}$ increases from zero to $\mathcal{I}_{max,S}$. Similarly, spending of error probabilities α_N and β_N for the test of non-inferiority follows functions $f_N(\mathcal{I})$ and $g_N(\mathcal{I})$ as $\mathcal{I}$ increases from zero to $\mathcal{I}_{max}$. At the design stage, we make a working assumption that a specific sequence of information levels will be observed and specify a combination of f_S, g_S, f_N, g_N, $\mathcal{I}_{max,S}$ and $\mathcal{I}_{max}$ for which boundaries converge to spend all four error probabilities exactly by the end of the study. We plan for K interim analyses at information levels

$$\mathcal{I}_k = k\,\mathcal{I}_{max,S}/K_S \quad \text{for } k = 1, \ldots, K_S \tag{2.7}$$

and

$$\mathcal{I}_k = \mathcal{I}_{max,S} + (k - K_S)(\mathcal{I}_{max} - \mathcal{I}_{max,S})/(K - K_S) \quad \text{for } k = K_S + 1, \ldots, K. \tag{2.8}$$

In practice, the test will adapt to observed information levels, maintaining type I error probabilities precisely but with small perturbations to the type II error rates.

Our choice of error spending functions is motivated by the cumulative error rates seen in optimal designs. Since these designs do not allow very early decisions of non-inferiority, we also delay spending α_N and β_S until information reaches a minimum threshold $\gamma\,\mathcal{I}_{max,S}$, where $0 \le \gamma \le 1$. This is a sensible feature since, with only a small amount of data, one cannot be confident that θ is both above $-\delta_N$ and below δ_S. We propose a family of designs with spending functions indexed by the parameter $\rho > 0$, similar in form to those for the two-decision problem used by Jennison and Turnbull (2003). The four error spending functions are:

$$f_N(\mathcal{I}) = \begin{cases} 0 & \text{if } \mathcal{I} < \gamma\,\mathcal{I}_{max,S} \\ \alpha_N\,(\mathcal{I}/\mathcal{I}_{max})^{\rho} & \text{if } \gamma\,\mathcal{I}_{max,S} \le \mathcal{I} < \mathcal{I}_{max} \\ \alpha_N & \text{if } \mathcal{I} \ge \mathcal{I}_{max} \end{cases}$$

$$f_S(\mathcal{I}) = \begin{cases} \alpha_S\,(\mathcal{I}/\mathcal{I}_{max,S})^{\rho} & \text{if } \mathcal{I} < \mathcal{I}_{max,S} \\ \alpha_S & \text{if } \mathcal{I} \ge \mathcal{I}_{max,S} \end{cases}$$

$$g_N(\mathcal{I}) = \begin{cases} \beta_N\,(\mathcal{I}/\mathcal{I}_{max})^{\rho} & \text{if } \mathcal{I} < \mathcal{I}_{max} \\ \beta_N & \text{if } \mathcal{I} \ge \mathcal{I}_{max} \end{cases}$$

$$g_S(\mathcal{I}) = \begin{cases} 0 & \text{if } \mathcal{I} < \gamma\,\mathcal{I}_{max,S} \\ \beta_S\,(\mathcal{I}/\mathcal{I}_{max,S})^{\rho} & \text{if } \gamma\,\mathcal{I}_{max,S} \le \mathcal{I} < \mathcal{I}_{max,S} \\ \beta_S & \text{if } \mathcal{I} \ge \mathcal{I}_{max,S}, \end{cases}$$

where $\gamma > 0$. Figure 2-6 shows these functions for the case $\rho = 1$ and $\gamma = 0.5$. When $\mathcal{I}_{Sf} < \mathcal{I}_{Nf}$, Brannath et al. (2003) comment on the desirability of spending the type I error probability α_S for the superiority objective more rapidly than that for the test of non-inferiority, α_N. This feature is built into our definitions of spending functions but there would be no difficulty in taking such considerations further and varying the values of ρ in the four spending functions.

Application of this error spending design with an observed sequence of information levels, $\mathcal{I}_1, \mathcal{I}_2, \ldots$, follows the general framework described by (Jennison and Turnbull, 2000, Chapter 7) for other types of error spending test. At the first few analyses with $\mathcal{I}_k < \gamma\,\mathcal{I}_{max,S}$ only the outer boundary values d_k and a_k are required. These are calculated to satisfy

$$P_{\theta=0}(\text{Declare "Superiority" by analysis } k) = f_S(\mathcal{I}_k) \tag{2.9}$$

and

$$P_{\theta=0}(\text{Declare "Inferiority" by analysis } k) = g_N(\mathcal{I}_k). \tag{2.10}$$

We do allow stopping to declare superiority when $f_N(\mathcal{I}_k) = 0$, even though this represents a type I error for the test of non-inferiority under $\theta = -\delta_N$. Similarly, we permit a decision of inferiority when $g_S(\mathcal{I}_k) = 0$, even though this is a type II error for the test of superiority under $\theta = \delta_S$. The probabilities of these outcomes are computed so they can be accounted for at later analyses when $f_N(\mathcal{I}_k)$ and $g_S(\mathcal{I}_k)$

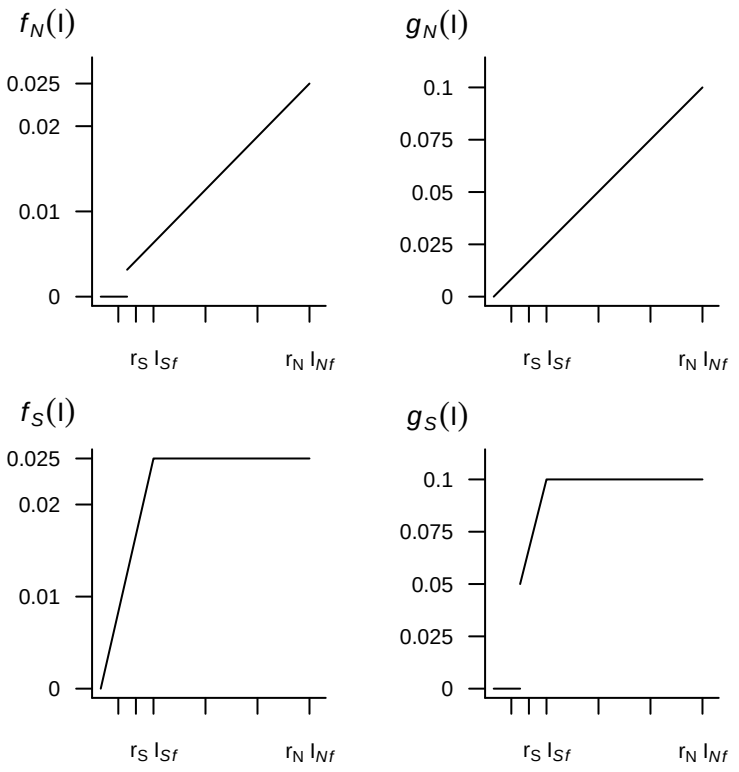

Figure 2-6: Error spending functions with $\rho = 1$ and $\gamma = 0.5$.

become positive.

For $\gamma \mathcal{I}_{max,S} \leq \mathcal{I}_k \leq \mathcal{I}_{max,S}$, we compute d_k and a_k to satisfy (2.9) and (2.10) and perform a two-dimensional search to find values b_k and c_k satisfying

$$P_{\theta=-\delta_N}(\text{Declare "Non-inferiority" or "Superiority" by analysis } k) = f_N(\mathcal{I}_k) \quad (2.11)$$

and

$$P_{\theta=\delta_S}(\text{Declare "Inferiority" or "Non-inferiority" by analysis } k) = g_S(\mathcal{I}_k). \quad (2.12)$$

Further details of how b_k and c_k can be found are given in Section 2.6.3.

At the first analysis $\tilde{K}_S$ where $\mathcal{I}_{\tilde{K}_S} \geq \mathcal{I}_{max,S}$ we calculate $d_{\tilde{K}_S}$ so that

$$P_{\theta=0}(\text{Declare "Superiority" by analysis } \tilde{K}_S) = \alpha_S$$

and set $c_{\tilde{K}_S} = d_{\tilde{K}_S}$. We find $a_{\tilde{K}_S}$ and $b_{\tilde{K}_S}$ satisfying (2.10) and (2.11) with $k = \tilde{K}_S$ and $\mathcal{I}_k = \mathcal{I}_{\tilde{K}_S}$. At subsequent analyses with $\mathcal{I}_k < \mathcal{I}_{max}$ we set $c_k = d_k = \infty$ and calculate a_k and b_k to satisfy (2.10) and (2.11). Finally, at the first analysis $\tilde{K}$ with $\mathcal{I}_{\tilde{K}} \geq \mathcal{I}_{max}$

we find $b_{\tilde{K}}$ satisfying

$$P_{\theta=-\delta_N}(\text{Declare "Non-inferiority" or "Superiority" by analysis } \tilde{K}) = \alpha_N$$

and set $a_{\tilde{K}} = b_{\tilde{K}}$.

By construction, this error spending procedure attains the type I error probabilities α_N and α_S exactly. The type II error probabilities may differ slightly from β_N and β_S but they will be close to these targets if the observed information levels are similar to the sequence defined by (2.7) and (2.8) which was assumed for planning purposes.

Suppose that α_S, α_N, β_N and β_S are specified, so $\mathcal{I}_{Sf}$ and $\mathcal{I}_{Nf}$ are fixed multiples of δ_S^{-2} and δ_N^{-2}, respectively. For given δ_N/δ_S, ρ, γ, K_S and K, and assuming analyses are scheduled according to (2.7) and (2.8), a two-dimensional search can be conducted to find the inflation factors r_S and r_N which give power $1-\beta_N$ at $\theta = 0$ and $1-\beta_S$ at $\theta = \delta_S$. Within the ρ family, increasing ρ reduces the rate at which error is spent, leading to smaller inflation factors. Thus, for given α_S, α_N, β_N, β_S, γ, K_S, K and $\delta_N/\delta_S < 1$, say, there is a one-to-one correspondence between ρ and r_N. While inflation factors do increase gradually with K, broadly speaking, setting $\rho = 3$ gives an inflation factor around $r_N = 1.05$ and wide outer boundaries similar to an O'Brien and Fleming (1979) test, whereas $\rho = 1$ yields an inflation factor around 1.2 or 1.25 and narrower boundaries, as in a Pocock (1977) test.

Comparing the ρ family error spending tests with designs optimised for $F^\star$, we have found the ρ family tests to be highly efficient and achieve values of $F^\star$ within a few percent of the minimum possible for a given inflation factor r_N. We conclude that error spending designs in the ρ family are both efficient and sufficiently flexible to handle unpredictable group sizes or information levels. These findings are in keeping with those of Barber and Jennison (2002) for one-sided error spending tests with similar spending functions.

As an illustration of the preceding remarks, Figure 2-7 shows the expected sample size function for the design with $\alpha_N = \alpha_S = 0.025$, $\beta_N = \beta_S = 0.1$, $\delta_N = 0.1$, $\delta_S = 0.2$, $\rho = 1$, $\gamma = 0.5$, $K_S = 3$ and $K = 6$, as well as that for the optimal design minimising $F^\star$ for the same problem and group sizes. It is evident that the error spending design is highly efficient across the range of θ values and, overall, it achieves a value of $F^\star$ within 2% of the optimum.

If we consider the same example, but vary ρ from 0.5 to 3, we obtain designs with inflation factors r_N ranging from around 1.5 to 1.05. Figure 2-8 shows values of $F^\star$ for these error spending designs plotted against the inflation factor r_N for each design. The slightly lower curve gives the value of $F^\star$ achieved by optimal designs for this criterion with the same group sizes, which is around 2 to 4 per cent smaller than that of the error spending design. The levelling off of $F^\star$ as ρ decreases and r_N increases indicates there is no advantage in taking r_N greater than around 1.2, which is attained by ρ of

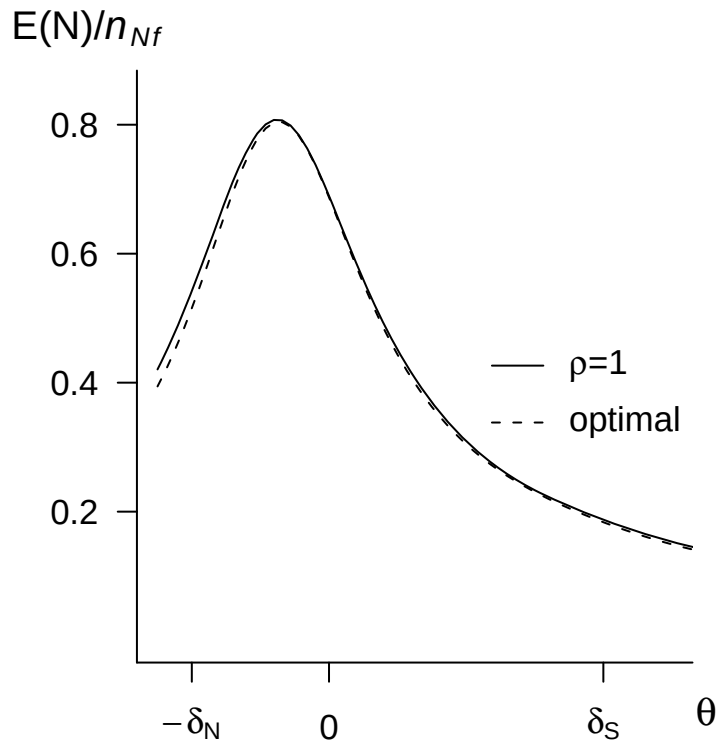

Figure 2-7: $E(N)/n_{Nf}$ for a 6-group error spending design with ρ=1 and $\gamma = 0.5$ and for the optimal 6-group design with analyses performed at the same information levels.

about 0.8.

We recommend 0.5 as a simple default value for γ. In a detailed assessment of a particular case one can go further and compare values of γ with respect to expected sample size functions under different sequences of information levels, paying particular attention to the effect of $\mathcal{I}_1$.

Another advantage of error spending tests is that they support use of the method of information monitoring, as proposed by Mehta and Tsiatis (2001). This approach can be used to manage a trial when the sample size needed for specific power depends on nuisance parameters which are only estimated once the trial is under way. One example of such a parameter is the variance of a normal response: thus, error spending and information monitoring provide a way to deal with unknown variance in the normal response problem introduced in Section 2.2.1.

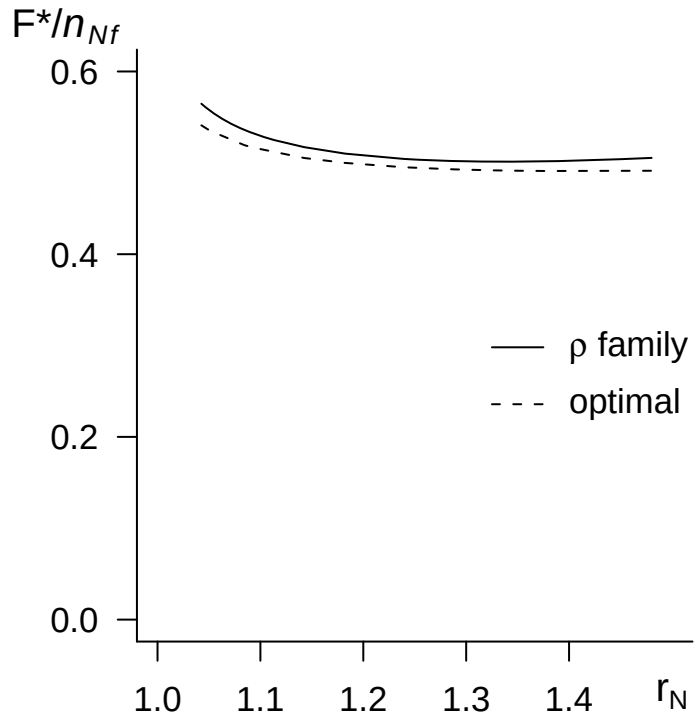

Figure 2-8: $F^\star/n_{Nf}$ for ρ family error spending designs with ρ in the range 0.5 to 3 and for optimal designs minimising $F^\star$ with the same sequences of information levels.

2.3 Optimal adaptive designs

2.3.1 Framework

The need for different sample sizes to test superiority and non-inferiority has prompted proposals for designs in which future group sizes are based on previously observed data. Such procedures are examples of adaptive group sequential designs, as proposed for one-sided tests by Schmitz (1993). These methods extend those of Section 2.2 by allowing each new group size to depend on previous data. Since this wider class includes non-adaptive group sequential designs as special cases, optimising over it yields a lower value of an objective function, such as $F^\star$, than that of the best non-adaptive design. We shall now explore the benefits of adaptation in reducing expected sample size in the three-decision problem and consider whether they justify the extra complexity of this approach. The definition of adaptive group sequential designs applies for general K but we shall focus on $K = 2$ for computational simplicity.

In the three-decision problem, adaptive designs may terminate at each analysis with a decision of inferiority, non-inferiority or superiority. However, if sampling continues at analysis k, the next group size is allowed to depend on Z_k. We seek the sequential

decision rule that minimises $F^\star$ subject to the error constraints (2.1) to (2.4). The optimal K-group adaptive design can be derived by the Lagrangian approach used in Section 2.2 for the non-adaptive case. The unconstrained problem is a Bayes decision problem, solvable by dynamic programming.

In solving the unconstrained problem for the case $K = 2$, suppose the first analyses takes place after n_1 observations. We approximate the continuous range of values for n_2 by giving M possible cumulative group sizes, $n_{2,1}, \ldots, n_{2,M}$, for the final analysis. We find the optimal critical values $b_{2,1}, \ldots, b_{2,M}$ and $d_{2,1}, \ldots, d_{2,M}$ to decide between inferiority, non-inferiority and superiority in each case. The task at the first analysis is to determine which of the following actions is optimal: stop for inferiority, stop for non-inferiority, stop for superiority, or continue to the final analysis with cumulative group size $n_{2,m}$ for a value of $m \in \{1, \ldots, M\}$. This gives the optimal design for a given value of n_1 and a further search over n_1 gives the two-stage adaptive design with the overall minimum of $F^\star$.

In numerical calculations we have used $M = 100$ and $n_{2,1}, \ldots, n_{2,M}$ equally spaced between n_1 and an upper limit $R\, n_{Nf}$ where $R > 1$. We performed sensitivity analyses to check there is no significant change to the design if M or R is increased. For most examples, we have found that when using $R = 1.35$, the optimal choice of n_2 is lower than $R\, n_{Nf}$ for all values of Z_1. We give further details of implementing the dynamic programming algorithm in Section 2.6.2.

2.3.2 Efficiency gains through adaptation

We illustrate with an example the possible efficiency gains from adaptation in two-stage designs. The results in Table 2.1 are for non-adaptive and adaptive two-stage designs which minimise $F^\star$ subject to error probabilities $\alpha_N = \alpha_S = 0.025$ and $\beta_N = \beta_S = 0.1$ for values of $\delta_{Sf}/\delta_{Nf} = (n_{Nf}/n_{Sf})^{1/2}$ ranging from 1 to $\sqrt{3}$. In the adaptive design the initial group size, n_1, is chosen optimally and the second group size, $n_2 - n_1$, is selected to be optimal for the observed Z_1. For the non-adaptive designs, n_1 and n_2 are fixed at optimal values for the objective function $F^\star$. The maximum value of n_2 in the adaptive designs ranges from $1.20\, n_{Nf}$ to $1.26\, n_{Nf}$ in the six cases of Table 2.1 whereas, in the non-adaptive designs, n_2 takes lower values between $1.12\, n_{Nf}$ and $1.17\, n_{Nf}$. In fact, for the case $n_{Nf}/n_{Sf} = 3$, values of n_1 greater than n_{Sf} help to minimise $F^\star$ in both the non-adaptive and adaptive designs, but lead to power for the test of superiority greater than the stipulated $1 - \beta_S = 0.9$. Reformulating this requirement as an inequality that power should be at least 0.9, leads to the designs reported here which have both higher power for the test of superiority (around 0.93) and lower $F^\star$.

The results in Table 2.1 show only minor benefits from adaptation. These benefits are greatest for intermediate values of the ratio n_{Nf}/n_{Sf} and in these cases there are areas of the adaptive design's continuation region at the first analysis where each of

$n_{Nf}/n_{Sf} =$	1.0	1.25	1.5	1.75	2.0	3.0
Optimal non-adaptive designs	81.8	75.3	71.7	69.1	67.0	64.6
Optimal adaptive designs	81.7	74.3	70.1	67.7	66.1	64.2

Table 2.1: Values of $100\,F^\star/n_{Nf}$ for optimal two-stage designs with error probabilities at most $\alpha_N = \alpha_S = 0.025$ and $\beta_N = \beta_S = 0.1$ for selected values of $n_{Nf}/n_{Sf} = (\delta_S/\delta_N)^2$.

the three final decisions is plausible. This leads to substantial variation of the optimal values for n_2 with Z_1, as displayed in Figure 2-9 for the case $n_{Nf}/n_{Sf} = 1.5$. In view of the variation in the optimal n_2, it is not surprising that the best non-adaptive design, with only a single value of n_2, is less efficient. On the other hand, Figure 2-9 suggests it might be sufficient to choose between just two sample sizes, $n_{N,2}$ and $n_{S,2}$ say, in the lower and upper continuation regions, either side of the "inner wedge". We refer to such a procedure as a restricted adaptive design.

We computed a two-stage restricted design minimising $F^\star$ for the case $n_{Nf}/n_{Sf} = 1.5$, with n_1 set at the value chosen for the unrestricted adaptive design. The dashed lines in Figure 2-9 show the continuation intervals, which differ slightly from the unrestricted design, and values of $n_{N,2}$ and $n_{S,2}$. In Figure 2-10, the expected sample size function for the restricted adaptive design lies very close to that of the optimal unrestricted adaptive design, demonstrating that the key improvement in the adaptive design comes from choosing a sample size appropriate to the most relevant decision choice, superiority vs non-inferiority or non-inferiority vs inferiority, and not from any further fine-tuning. Values of $100F^\star/n_{Nf}$ are 71.7 for the two-group non-adaptive test, 70.4 for the restricted adaptive test, and 70.1 for the two-group adaptive test.

Figure 2-10 also shows the expected sample size function for the optimal non-adaptive three-group design with $n_{Nf}/n_{Sf} = 1.5$ and cumulative sample sizes equal to the values of n_1, $n_{S,2}$ and $n_{N,2}$ in the restricted adaptive design. Since the third analysis is only used to distinguish between inferiority and non-inferiority, this is an example from our class of non-adaptive designs with $K = 3$ and $K_S = 2$ (and $c_3 = d_3 = \infty$). This three-group non-adaptive design has lower expected sample size than the optimal adaptive two-stage design across the range of θ values and it is significantly more efficient at low values of θ; its value of $100F^\star/n_{Nf}$ is 66.8, compared to 70.1 for the optimal adaptive two-stage design. Our conclusions here concur with those of Jennison and Turnbull (2006a) about the two-decision problem: while adaptivity can lead to a small increase in efficiency, similar or larger improvements can be achieved with one additional interim analyses in a non-adaptive group sequential design. In view of the minor benefits accruing from adaptive choice of group size in the case $K = 2$, we have

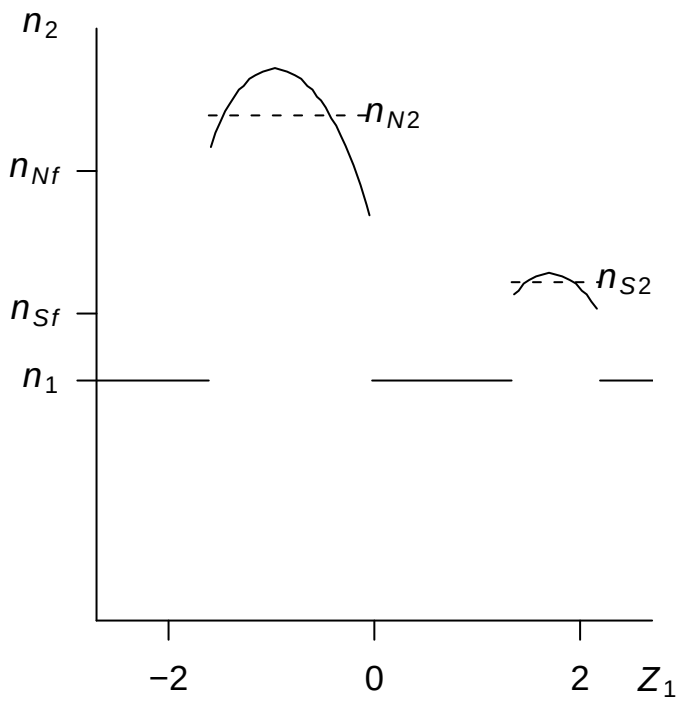

Figure 2-9: Final sample size, n_2, as a function of Z_1 for the optimal adaptive design for $n_{Nf}/n_{Sf} = 1.5$ (solid lines) and sample sizes $n_{N,2}$ and $n_{S,2}$ for the optimal restricted adaptive design (dashed lines).

not carried out computation of optimal adaptive designs for higher values of K.

2.3.3 Competing adaptive methods

The sample size function for the optimal adaptive design in Figure 2-9 is qualitatively different from that arising from a conditional power rule, where sample size rises as Z_1 decreases, at least within each continuation interval. While optimal adaptive designs offer modest gains over their non-adaptive counterparts, the following comparisons show published adaptive methods with sub-optimal sampling rules can be less efficient than simpler non-adaptive designs.

Koyama et al. (2005) propose an adaptive two-stage procedure for simultaneous testing of superiority and non-inferiority. After the first stage, stopping is possible for inferiority, non-inferiority or superiority. If the trial continues, the second stage sample size is set as a function of the first stage test statistic, Z_1. The sample size function and terminal decision rules are chosen to achieve specified overall error probabilities α_N, α_S, β_N, and β_S. Koyama et al. (2005) provide an example with $\delta_N = 1.0$, $\delta_S = 0.5$, $\sigma = 4$, $\alpha_N{=}\alpha_S = 0.025$, $\beta_N = 0.1$ and $\beta_S = 0.2$. While we have focused on designs

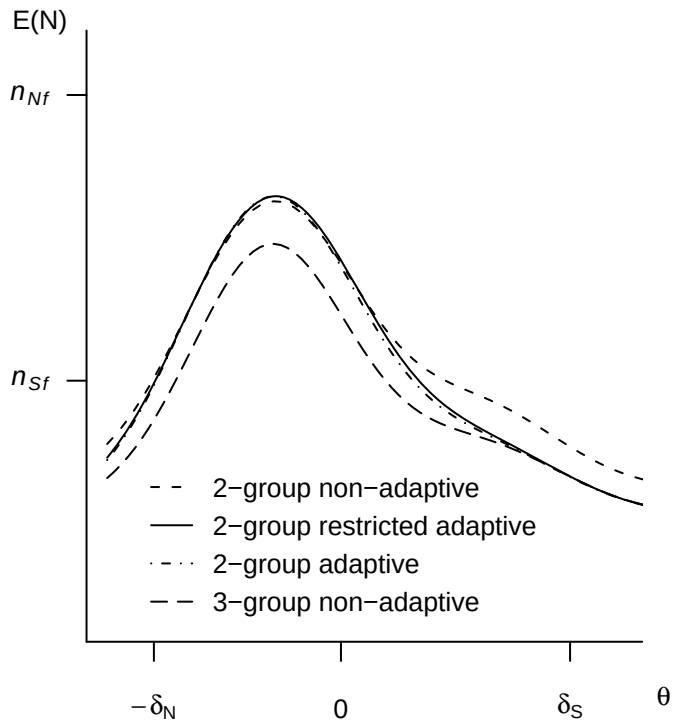

Figure 2-10: Expected sample size functions for optimal non-adaptive, restricted adaptive and adaptive 2-group designs and the optimal non-adaptive 3-group design.

with $\delta_S > \delta_N$, our framework also applies to the case $\delta_S < \delta_N$ studied by Koyama et al. (2005), Lai et al. (2006) and Brannath et al. (2003). We have compared the adaptive procedure of Koyama et al. with a two-stage non-adaptive design with the same error probabilities optimised for $F^\star$. In this design, the first analysis is scheduled after 337 observations per treatment and the final analysis after 1200 observations. Expected sample sizes per treatment are shown in Table 2.2. Not only is the non-adaptive procedure more efficient, but its maximal sample size per treatment is only 1200, compared to more than 1500 for the adaptive design.

The method proposed by Shih et al. (2004) falls within the framework for 2-stage adaptive procedures defined in Section 2.3.1. Early stopping for futility, non-inferiority or superiority is possible at the first analysis and critical values at both analyses are chosen to control the overall type I error probabilities at specified values α_N and α_S. The procedure does not aim to achieve a particular overall power, rather the second stage sample size is chosen with reference to conditional power given the first stage data. We have simulated the design presented for the survival data example in Section 3 of Shih et al. (2004) and found the overall power curves and expected sample size function of this design: with a survival endpoint, "sample size" should be interpreted as the

θ	$E_\theta(N)$ for Koyama et al's design	$E_\theta(N)$ for a 2-group non-adaptive design
$-\delta_N$	337	337
0	643	625
δ_S	996	937

Table 2.2: Comparison between the adaptive design of Koyama et al. and an optimal non-adaptive design.

number of events observed at termination. We constructed a non-adaptive 2-group sequential design with the same type I error probability and overall power curves at least as high as those of the procedure of Shih et al. (2004), over the range of effect sizes. This non-adaptive group sequential design had lower expected sample size by between 3% and 11% at values of θ in the range $-\delta_N$ to $2\,\delta_N$. We attribute the lower efficiency of the adaptive procedure to the choice of sample size function: for the optimal adaptive rule, values of n_2 are highest in the middle of each arm of the continuation region and lower nearer the boundary points, whereas the conditional power construction implies n_2 increases monotonically as $\hat{\theta}$ decreases.

Wang et al. (2001) propose an adaptive group sequential closed (AGSC) procedure which starts out as a group sequential design for testing superiority, but can shift adaptively between superiority and non-inferiority objectives. When $\delta_N < \delta_S$ and $n_{Nf} > n_{Sf}$, the initial design has n_{Sf} observations and K analyses. At each interim analysis, conditional power calculations determine whether to shift to the non-inferiority objective. If so, group sizes are increased to lead to a final sample size of n_{Nf} at analysis K with down-weighting as in the method of Cui et al. (1999) to maintain the type I error rate. Type II error rates are not controlled directly but are governed by n_{Nf}, n_{Sf}, the group sequential stopping boundary and the adaptive decision rule.

We evaluated the AGSC method by simulation with one million replicates. We assumed normal responses with $\sigma^2 = 9$, $\alpha_N = \alpha_S = 0.025$, $\delta_N = 0.4$ and $\delta_S = 0.8$. The initial design had five equally spaced analyses and a total of $n_{Sf} = 221$ observations per treatment arm, increasing to $n_{Nf} = 883$ under adaptation. Figure 2-11 compares the AGSC method and a non-adaptive 5-group sequential design with $K_S = 2$, $n_{K_S} = 221$, $K = 5$, and $n_5 = 883$. The non-adaptive design has higher power and a substantially lower expected sample size function. Since the AGSC method has no lower boundary to allow stopping for inferiority, its high expected sample size under low values of θ is to be expected. At higher effect sizes, using non-sufficient statistics as a result of down-weighting later observations is a source of inefficiency. More important, we believe, is the reliance on uncertain estimates of $\hat{\theta}$ at the interim analyses in making the decision

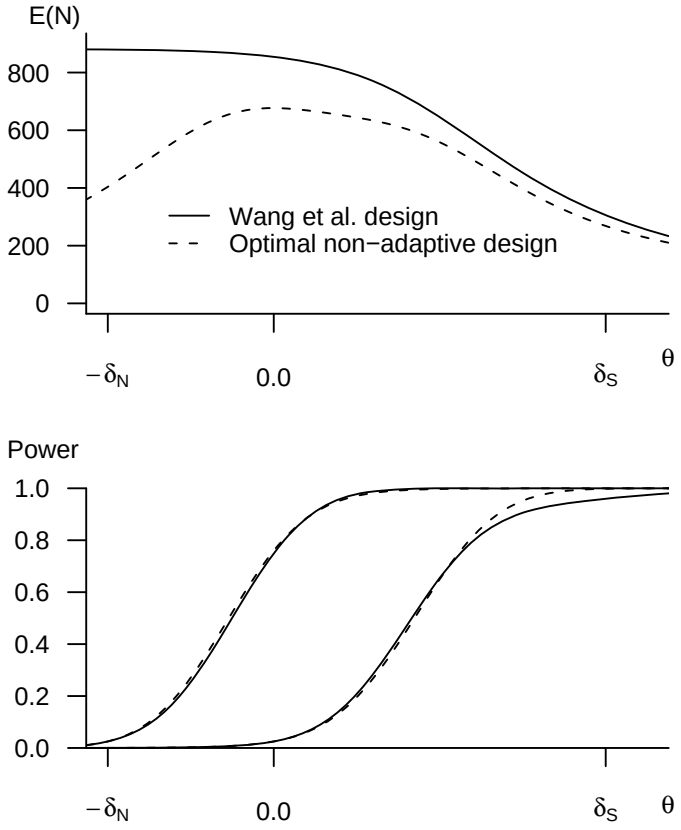

Figure 2-11: Expected sample size functions and power for non-inferiority and superiority, for the AGSC design of Wang et al. (solid line) and an optimal non-adaptive 5-group sequential design (dashed line).

to increase sample size four-fold. While we have found the addition of a lower futility boundary to the AGSC method improves performance for low values of θ, the method still fails to match the performance of the non-adaptive group sequential test at higher effect sizes.

2.4 An example in type 2 diabetes

We shall illustrate how the method can be applied, to a clinical trial in type 2 diabetes. EMEA (2002) recommends decrease from baseline HbA1c, a measure of blood glucose control, as the primary endpoint for studies of type 2 diabetes. In the trial reported by Home et al. (2007), the response was the percentage decrease in HbA1c, the non-inferiority margin was $\delta_N = 0.4$ and a standard deviation of 1.4 was used in the sample size calculation. Göke et al. (2007) report a clinical trial with power to detect an improvement of $\delta_S = 0.5$ under the new treatment and, again, a standard deviation of

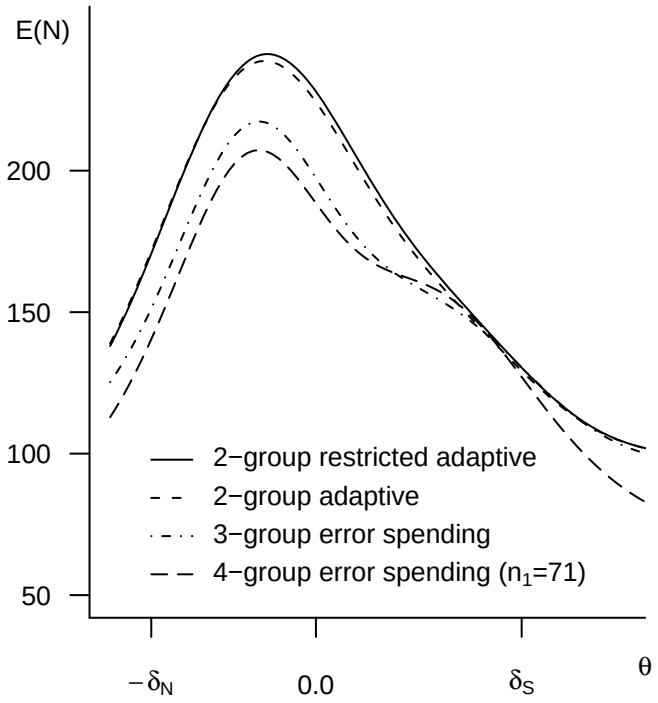

Figure 2-12: Expected sample size functions for designs in the type 2 diabetes example.

1.4 was used to determine sample size.

Consider designing a trial to compare a new treatment for type 2 diabetes against a standard, testing for both superiority and non-inferiority. Suppose responses are normally distributed with $X_{Ai} \sim N(\mu_A, \sigma^2)$ on the new treatment and $X_{Bi} \sim N(\mu_B, \sigma^2)$ on the standard. Denoting the treatment effect by θ=$\mu_A - \mu_B$, we wish to test simultaneously the null hypothesis $H_{N,0}$: $\theta \leq -0.4$ against $\theta > -0.4$ with power specified at $\theta = 0$ and the null hypothesis $H_{S,0} : \theta \leq 0$ against $\theta > 0$ with power at $\theta = 0.5$. Thus, we set $\delta_N = 0.4$, $\delta_S = 0.5$, $\alpha_N = 0.025$, $\alpha_S = 0.025$, $\beta_N = 0.1$, and $\beta_S = 0.1$ in our general framework. With $\sigma = 1.4$, fixed sample sizes per treatment are $n_{Sf} = 165$ and $n_{Nf} = 258$ for the two individual hypothesis tests.

For n_{Nf}/n_{Sf} around 1.5, the results of Section 2.3.2 indicate adaptation may be helpful if only two analyses are possible. We computed an adaptive two-group design optimised for $F^\star$ with $n_1 = 99$ and no upper limit for the second group size. We also derived a "restricted adaptive" design, as introduced in Section 2.3.2, where $n_1 = 99$ and, if sampling continues, the choice of the final sample size is either 198 or 309. The two upper curves in Figure 2-12 are the expected sample size functions for these restricted adaptive and adaptive designs. We see that restricting the second group size to just two values has a negligible effect on efficiency.

Comparison of expected sample size functions allows an informed choice of a suitable design. In making this choice, investigators may also consider the logistical challenges of setting up a trial with data-driven choice of the second group size. Information leakage should be considered since, in both the adaptive and restricted adaptive designs, knowledge of the second stage sample size provides an indication of the first stage results. In this joint testing problem, leakage can also be an issue for a non-adaptive group sequential design: in a 3-group procedure with $K_S = 2$, continuation past the second analysis implies the new treatment has not been found to be superior and the decision will be either "non-inferior" or "inferior".

The error spending method of Section 2.2.5 can be used to give a design with close to optimal efficiency as well as the flexibility to deal with unpredictable group sizes. If we use ρ family error spending functions with $\rho = 1$ and design for $K = 3$ analyses with $K_S = 2$ and $\gamma = 0.4$, the inflation factors are $r_S = 1.167$ and $r_N = 1.195$, so $n_{max,S} = 1.167 n_{Sf} = 193$ and the maximum sample size is $n_{max} = 1.195\, n_{Nf} = 308$. If observed sample sizes follow the design pattern of $n_1 = 97$, $n_2 = 193$ and $n_3 = 308$, power of 0.9 is attained exactly in both hypothesis tests. The expected sample size function for this design shown in Figure 2-12 is almost identical to that obtained by a 3-group sequential design with $r_S = r_N = 1.2$ optimised for $F^\star$.

Suppose patient accrual is lower than expected and only $\tilde{n}_1 = 71$ responses are observed at the first analysis. Since $\tilde{n}_1 < \gamma n_{max,S}$, there is no inner wedge at the first analysis. If accrual remains slow throughout the trial, a fourth analysis will be needed to reach n_{max} but the error spending design adjusts easily to the new sequence of sample sizes. Suppose we observe $\tilde{n}_2 = 144$, $\tilde{n}_3 = 220$ and $\tilde{n}_4 = 308$. The critical values for what is now a 4-group design are computed following the prescription in Section 2.2.5: the resulting boundaries are shown in Figure 2-13. The type I error probabilities are automatically controlled at $\alpha_N = 0.025$ and $\alpha_S = 0.025$ and the attained type II error probabilities are $P_{\theta=0}(\text{Conclude "Inferiority"}) = 0.102$ and $P_{\theta=\delta_S}(\text{Conclude "Inferiority or Non-inferiority"}) = 0.088$, both close to their intended values of $\beta_N = \beta_S = 0.1$. The inner wedge plays an important role, allowing stopping for any of the three possible outcomes, superiority, non-inferiority and inferiority, at analyses two and three. The expected sample size function in this case is the lowest curve in Figure 2-12. So, not only does the error spending design deal well with the observed pattern of group sizes, but results for this four group design show it gains efficiency by adapting to a higher number of smaller group sizes when these arise in practice.

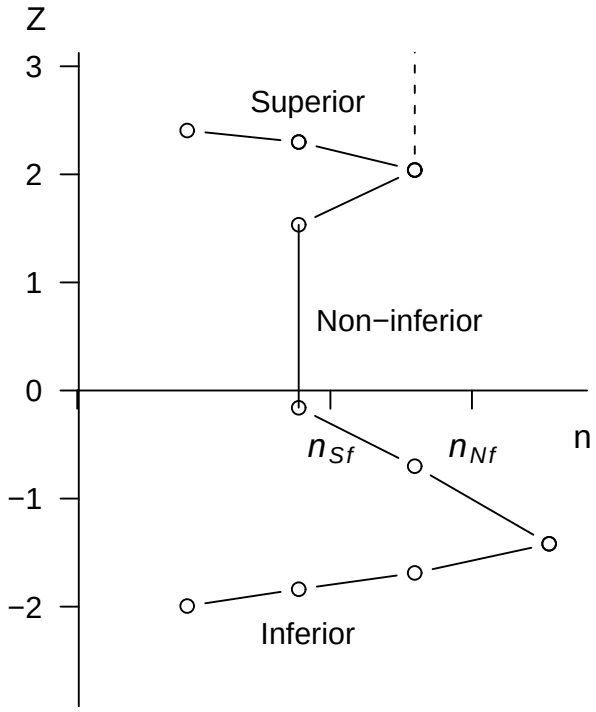

Figure 2-13: Critical values for 4-group error spending design.

2.5 Discussion

We have introduced a framework to define group sequential designs which test simultaneously for superiority and non-inferiority and allow early stopping for any one of the three conclusions of superiority, non-inferiority and inferiority. We can compute the design of this type which minimises a weighted combination of expected sample sizes at several effect sizes. We have also defined error spending versions of these designs which can handle unpredictable group sizes while retaining almost optimal efficiency. Expressing these error spending designs in terms of information for the effect size parameter shows they are applicable to a wide variety of response types and can deal with nuisance parameters governing the sample size needed for a specific power through the information monitoring approach of Mehta and Tsiatis (2001).

We have followed other authors in addressing the situation where the non-inferiority margin δ_N is smaller than the effect size δ_S at which power is set in the test for superiority. Here, much larger sample sizes are required when the study focuses on distinguishing between non-inferiority and inferiority. If such large sample sizes are known to be available if needed, one might expect investigators to consider increasing the power of the test for superiority to detect smaller effect sizes than δ_S. If this occurs, n_{Sf} will increase and the ratio n_{Nf}/n_{Sf} will be closer to one. Our framework

will still be appropriate and the inner wedge, which allows early stopping to declare non-inferiority, will play an important role in such cases.

In the two-decision problem, Barber and Jennison (2002) found the greatest benefit of an additional interim analyses to arise when moving from a fixed sample test to a two-group test. For the three-decision problem, two analyses are required simply to meet the different error constraints for the pair of hypothesis tests. When sample sizes for the two testing objectives are very different, a total of four analyses is needed to allow two suitably placed analyses for each hypothesis test. Thus, it is a feature of the group sequential designs for the three decision problem that a larger number of interim analyses is likely to be worthwhile than for group sequential designs for the two-decision problem.

In the "adaptive" designs considered in Section 2.3, future group sizes are based on current data, in particular the observed effect $\hat{\theta}$. Remember, though, that our "non-adaptive" group sequential designs also respond to the observed data: the stopping rule provides a very definite response and, when $n_{Nf} > n_{Sf}$, the absence of an upper continuation region at the last few analyses shows a shift of focus to the test between non-inferiority and inferiority.

In exploring the adaptive choice of group sizes, we have found only minor benefits of adaptation in two-stage designs, the case most often considered in the literature. In fact, we saw in Section 2.3.3 that non-adaptive group sequential designs can outperform some proposed adaptive methods with the same number of analyses. The greatest advantage we have found of an adaptive over an optimal non-adaptive 2-group design is around 3% of the fixed sample size. This may be a significant benefit in a clinical trial with thousands of patients — but then there is reason to pursue the even greater benefits of a non-adaptive 3-group sequential design. We have not invested effort in deriving optimal adaptive designs with three or more analyses as we do not anticipate substantively different results from the two-group case.

2.6 Proofs and derivations

2.6.1 Proof of monotonicity of type I and type II error probabilities

It seems intuitive that the probability of rejecting a null hypothesis such as $H_{S,0}$: $\theta \leq 0$ should increase with θ in any sensible experimental design. A coupling argument can provide a proof for some group sequential designs (see, for example, Jennison and Turnbull (2000, Page 183)), but this approach does not extend to group sequential designs with an inner wedge. Adaptive designs pose further problems, indeed Jennison and Turnbull (2003, Section 4.2) present an adaptive design with a non-monotone power function. However, Shih et al. (2004) are able to prove monotonicity of the type I error probability within the null hypothesis for two-stage adaptive designs. We

now generalise their result to K-group designs.

Consider first the non-adaptive case and a K-group design, as defined in Section 2.2.1. Let $f_k(z_k;\theta)$ denote the density of Z_k at analysis k under treatment effect θ in the absence of any prior early stopping. Define $p_{kc}(z_k)$ to be the conditional probability that $Z_1,\ldots,Z_{k-1}$ lie in the continuation regions

$$(a_1,b_1)\cup(c_1,d_1),\ldots,(a_{k-1},b_{k-1})\cup(c_{k-1},d_{k-1}),$$

given that $Z_k=z_k$. Since Z_k is sufficient for θ, this probability does not depend on θ. We can write

$$P_\theta(\text{Declare "Superiority"})=\sum_{k=1}^{K}\int_{d_k}^{\infty} f_k(z_k;\theta)\,p_{kc}(z_k)\,dz_k. \tag{2.13}$$

Now, marginally, $Z_k\sim N(\theta\sqrt{\mathcal{I}_k},\,1)$. Thus, if $d_k\geq 0$, $f_k(z_k;\theta)$ is an increasing function of θ for all $\theta\leq 0$ and $z_k>d_k$. It follows that all the integrands in the right hand side of (2.13) are increasing in θ for $\theta\leq 0$. Hence, as long as $d_k\geq 0$ for each $k=1,\ldots,K$, P_θ(Declare "Superiority") increases monotonely with θ for $\theta\leq 0$ and the maximum type I error probability over $\theta\leq 0$ occurs at $\theta=0$. The condition $d_k\geq 0$ implies that stopping to reject $H_{S,0}$: $\theta\leq 0$ is only possible when $\hat{\theta}_k\geq 0$, which is to be expected in any sensible design.

A similar argument shows P_θ(Declare "Non-inferiority or Superiority") is monotone over $\theta\leq-\delta_N$. In this case, the integrals in (2.13) have range $(b_k,c_k)\cup(d_k,\infty)$ and the condition for integrands to be increasing for $\theta\leq-\delta_N$ becomes $b_k\geq-\delta_N\sqrt{\mathcal{I}_k}$, so $H_{N,0}$: $\theta\leq-\delta_N$ is only rejected when $\hat{\theta}_k\geq-\delta_N$. The same approach can be used to establish monotonicity results for type II error probabilities:

$$P_\theta(\text{Conclude "Inferiority"})$$

decreases with θ for $\theta\geq 0$ as long as this decision only occurs when $\hat{\theta}_k\leq 0$, and

$$P_\theta(\text{Conclude "Inferiority" or "Non-inferiority"})$$

decreases with θ for $\theta\geq\delta_S$ as long as this decision requires $\hat{\theta}_k\leq\delta_S$.

We can obtain results for adaptive group sequential designs by essentially the same argument. Since the sample size at each analysis now depends on previous responses, the sum over k in (2.13) becomes a double sum over k and the set of possible sequences $\{\mathcal{I}_1,\ldots,\mathcal{I}_k\}$. In some designs the critical value d_k may depend on the whole sequence $\{\mathcal{I}_1,\ldots,\mathcal{I}_k\}$. It is useful, conceptually, to define the sequence of Z-statistics at all information levels that might arise, noting the joint distribution theory stated at the start of Section 2.2.5 applies to this whole sequence. We let $f_k(z_k,\mathcal{I}_k;\theta)$ denote the

$N(\theta\sqrt{\mathcal{I}_k},\, 1)$ density of Z_k for treatment effect θ and information level $\mathcal{I}_k$ in the absence of any prior early stopping. We define

$$p_{kc}(z_k, \mathcal{I}_1, \ldots, \mathcal{I}_k)$$

to be the conditional probability of following the sequence of information levels $\mathcal{I}_1, \ldots, \mathcal{I}_k$ to reach analysis k with information $\mathcal{I}_k$ and $Z_k = z_k$, given that Z_k takes this value when information is equal to $\mathcal{I}_k$. Again, this conditional probability does not depend on θ. In place of (2.13) we now have

$$P_\theta(\text{“Superiority”}) = \sum_{k=1}^{K} \sum_{\{\mathcal{I}_1,\ldots,\mathcal{I}_k\}} \int_{d_k(\mathcal{I}_1,\ldots,\mathcal{I}_k)}^{\infty} f_k(z_k, \mathcal{I}_k; \theta)\, p_{kc}(z_k, \mathcal{I}_1, \ldots, \mathcal{I}_k)\, dz_k.$$

As before, all the integrands in this equation are monotone increasing in θ, as long as each critical value $d_k(\mathcal{I}_1, \ldots, \mathcal{I}_k)$ is positive and, hence, the maximum type I error rate over $\theta \leq 0$ occurs at $\theta = 0$. Results for other error probabilities follow as before with the same conditions on critical values when these are expressed in terms of the final $\hat{\theta}_k$.

2.6.2 Derivation of optimal group sequential designs by solving Bayes decision problem

We illustrate our methods in the derivation of a design minimising $F^\star$ subject to error constraints (2.1) to (2.4). In this case, we place a five point prior distribution on θ with probability 1/5 at $-\delta_N$, $-\delta_N/2$, 0, $\delta_S/2$ and δ_S. We define a loss function associated with decisions on termination D_I: declare inferiority, D_N: declare-inferiority, and D_S: declare superiority,

$$L(D, \theta) = \begin{cases} k_1 & \text{for } D = D_N \text{ or } D_S \text{ and } \theta = -\delta_N \\ k_2 & \text{for } D = D_S \text{ and } \theta = 0 \\ k_3 & \text{for } D = D_I \text{ and } \theta = 0 \\ k_4 & \text{for } D = D_I \text{ or } D_N \text{ and } \theta = \delta_S \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

With a cost $c(\theta) = 1$ per observation at each value of θ, the total expected cost is

$$F^\star + \{k_1\, P_{\theta=-\delta_N}(D_N \cup D_S) + k_2\, P_{\theta=0}(D_S) + k_3\, P_{\theta=0}(D_I) + k_4\, P_{\theta=\delta_S}(D_I \cup D_N)\}/5.$$

We use dynamic programming, as described below, to solve this unconstrained Bayes decision problem for either non-adaptive or adaptive designs. It then remains to perform a numerical search for values of k_1, k_2, k_3 and k_4 which give a solution satisfying the error probability constraints (2.1) to (2.4). The standard Lagrangian

argument implies that this Bayes sequential decision rule minimises $F^\star$ among all designs satisfying (2.1) to (2.4).

Consider first the non-adaptive case where n_k, $k = 1, \ldots, K$, are pre-specified. Let $\pi^{(k)}(\theta|z_k)$ denote the posterior distribution of θ given $Z_k = z_k$. If sampling continues until the final analysis K, a decision D_I, D_N, or D_S must be chosen. The critical values at this analysis are obtained by solving $k_3\,\pi^{(K)}(0|z_K) = k_1\,\pi^{(K)}(-\delta_N|z_K)$ to find b_K, and $k_4\,\pi^{(K)}(\delta_S|z_K) = k_2\,\pi^{(K)}(0|z_K)$ to find d_K; the monotone likelihood ratio property of the normal distribution implies each of these equations has a unique solution. The dynamic programming algorithm works backwards from this point to find the optimal decision rule at earlier analyses.

If $Z_k = z_k$, the expected loss when stopping to make the Bayes optimal decision at stage k is

$$\gamma^{(k)}(z_k) = \min\{k_3\,\pi^{(k)}(0|z_k) + k_4\,\pi^{(k)}(\delta_S|z_k),\; k_1\,\pi^{(k)}(-\delta_N|z_k) + k_4\,\pi^{(k)}(\delta_S|z_k),$$

$$k_1\,\pi^{(k)}(-\delta_N|z_k) + k_2\,\pi^{(k)}(0|z_k)\},$$

the minimum of the expected costs of stopping for inferiority, non-inferiority or superiority.

We denote by $F^{k+1}(z_{k+1}|z_k)$ the conditional cumulative distribution function of Z_{k+1} given $Z_k = z_k$. For $k = 1, \ldots, K-2$, the additional expected cost for proceeding from stage k to stage $k+1$ and acting optimally thereafter is

$$\begin{aligned}\beta^{(k)}(z_k) &= (n_{k+1} - n_k)\sum_{i=1}^{5} c(\theta_i)\pi^{(k)}(\theta_i|z_k) \\ &+ \int \min\{\beta^{(k+1)}(z_{k+1}), \gamma^{(k+1)}(z_{k+1})\}dF^{(k+1)}(z_{k+1}|z_k) \qquad (2.15)\end{aligned}$$

while at stage $K-1$, we have

$$\begin{aligned}\beta^{(K-1)}(z_{K-1}) &= (n_K - n_{K-1})\sum_{i=1}^{5} c(\theta_i)\,\pi^{(K-1)}(\theta_i|z_{K-1}) \\ &+ \int \gamma^K(z_K)dF^{(K)}(z_K|z_{K-1}). \qquad (2.16)\end{aligned}$$

The functions $\beta^{(k)}(z_k)$ can be calculated recursively, working backwards from analysis $K-1$: using the stage $k+1$ stopping boundary and values of $\beta^{(k+1)}$ and $\gamma^{(k+1)}$ previously calculated on a grid of z_{k+1} values, the integral in (2.15) can be found by numerical integration using, say, Simpson's rule. At each analysis k, the roots of $\beta^{(k)}(z_k) = \gamma^{(k)}(z_k)$ are found by a numerical search and these roots define the stopping boundaries.

The above method can be extended to find optimal adaptive designs using the

approach followed by Jennison and Turnbull (2006a) for the two-decision problem. Consider the case $K = 2$, with n_1 fixed and n_2 allowed to take values in the set $\{n_{2,1}, \ldots, n_{2,M}\}$. We first find the M pairs of critical values $b_{2,m}$ and $d_{2,m}$ defining the optimal decisions when the second analysis takes place at cumulative sample size $n_{2,m}$, $m = 1, \ldots, M$. We then divide the range of values of Z_1 into intervals over which each of the following actions is found to be optimal: stop and declare inferiority, stop and declare non-inferiority, stop and declare superiority, continue to analysis 2 with cumulative group size $n_{2,m}$, $m = 1, \ldots, M$. As before, a numerical search is performed to find the set of costs k_1, k_2, k_3, and k_4 for which the solution satisfies the error probability constraints (2.1) to (2.4) and this gives the solution to the original constrained problem. This process is then nested within a search over n_1 to optimise both group sizes.

2.6.3 Calculation of critical values for error spending designs

Consider an analysis k with $\gamma\mathcal{I}_{max,S} \leq \mathcal{I}_k \leq \mathcal{I}_{max,S}$, the case where all four critical values, a_k, b_k, c_k and d_k, are required. We assume boundary values for analyses 1 to $k-1$ have already been calculated. Define the increments in error probabilities under $\theta = 0$ for analysis k

$$\Delta f_S^k = f_S(\mathcal{I}_k) - f_S(\mathcal{I}_{k-1}) \quad \text{and} \quad \Delta g_N^k = g_N(\mathcal{I}_k) - g_N(\mathcal{I}_{k-1}).$$

For the other two error probabilities, under $\theta = -\delta_N$ and δ_S, we set increments

$$\Delta f_N^k = f_N(\mathcal{I}_k) - f_N(\mathcal{I}_{k-1}) \quad \text{and} \quad \Delta g_S^k = g_S(\mathcal{I}_k) - g_S(\mathcal{I}_{k-1})$$

unless this is the first analysis with $\mathcal{I}_k \geq \gamma\mathcal{I}_{max,S}$, in which case we take

$$\Delta f_N^k = f_N(\mathcal{I}_k) - P_{\theta=-\delta_N}(\text{Stop to declare superiority by analysis } k-1)$$

and

$$\Delta g_S^k = g_S(\mathcal{I}_k) - P_{\theta=\delta_S}(\text{Stop to declare inferiority by analysis } k-1)$$

to account for the error probability incurred at analyses where $f_N(\mathcal{I})$ and $g_S(\mathcal{I})$ are zero.

We denote the continuation region at analysis i by $\mathcal{C}_i = (a_i, b_i) \cup (c_i, d_i)$. Two one-dimensional searches can be used to find a_k and d_k satisfying

$$P_{\theta=0}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \leq a_k) = \Delta g_N^k$$

and

$$P_{\theta=0}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \geq d_k) = \Delta f_S^k.$$

Let

$$\Delta f_N^{k1} = P_{\theta=-\delta_N}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \geq d_k)$$

and

$$\Delta g_S^{k1} = P_{\theta=\delta_S}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \leq a_k).$$

We now want to find b_k and c_k satisfying

$$P_{\theta=-\delta_N}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k, c_k]) = \Delta f_N^k - \Delta f_N^{k1} \tag{2.17}$$

and

$$P_{\theta=\delta_S}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k, c_k]) = \Delta g_S^k - \Delta g_S^{k1}. \tag{2.18}$$

Since b_k and c_k must lie in the interval $[a_k, d_k]$, an upper bound b_k^u for b_k is found by solving

$$P_{\theta=-\delta_N}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k^u, d_k]) = \Delta f_N^k - \Delta f_N^{k1} \tag{2.19}$$

and a lower bound c_k^l for c_k by solving

$$P_{\theta=\delta_S}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [a_k, c_k^l]) = \Delta g_S^k - \Delta g_S^{k1}. \tag{2.20}$$

Using these values of b_k^u and c_k^l, we can now find a lower bound b_k^l for b_k as the solution to

$$P_{\theta=-\delta_N}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k^l, c_k^l]) = \Delta f_N^k - \Delta f_N^{k1}$$

and an upper bound c_k^u for c_k as the solution to

$$P_{\theta=\delta_S}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1},\, Z_k \in [b_k^u, c_k^u]) = \Delta g_S^k - \Delta g_S^{k1}.$$

We have now reduced the original interval $[a_k, d_k]$ to $[b_k^l, c_k^u]$ and can repeat the same steps with c_k^u in place of d_k in (2.19) and b_k^l in place of a_k in (2.20). We have found repeated iterations of these steps to give an efficient method for finding b_k and c_k satisfying (2.17) and (2.18).

CHAPTER 3

Control of type I error when applying the CRP principle in an error spending design

3.1 Introduction

3.1.1 The CRP principle for known future information sequences

In the previous chapter we considered non-adaptive and adaptive group sequential methods for simultaneous testing of superiority and non-inferiority. We shall now consider another adaptive method that can be used within a group sequential framework. There has recently been an increased interest in adaptive methods that allow for changes to the design of an on-going clinical trial. While sample size modification based on the observed treatment effect has attracted a lot of attention, other types of design changes may also be of importance. It may for example be of interest to adapt the design to external information that was not available prior to the start of the trial. Withdrawal of a competing drug from the market may result in interest in detecting a smaller effect size than when the original design was planned.

Since both classical group sequential designs and adaptive designs have many useful properties, it would be very appealing to be able to combine the advantages of these approaches. In particular, an approach that enjoys all the benefits of error spending designs and in addition has the flexibility of adaptive methods would be of great value. One interesting proposal is due to the conditional rejection probability (CRP) principle proposed by Müller and Schäfer (2001) and later generalised by Müller and Schäfer (2004). In the CRP principle it is possible to re-design the remainder of the trial by calculating the conditional rejection probability under H_0, that is, the probability under H_0 of rejecting the null hypothesis given the data observed so far. This conditional rejection probability can then be used to design a new trial, with type I error equal to the conditional rejection probability. We shall now introduce how the method can be

applied through an example.

Suppose that the observations in treatment groups A and B are independent and normally distributed with $X_{Aj} \sim N(\mu_A, \sigma^2)$ and $X_{Bj} \sim N(\mu_B, \sigma^2)$ $j = 1, 2, \ldots,$ where the common variance σ^2 is known. The two treatment groups are compared in a clinical trial that is monitored through the error spending method, following the general framework described in Section 1.2.5. The objective of the trial is to make inference about the parameter $\theta{=}\mu_B - \mu_A$. A one-sided group sequential design, of the type without futility boundary described in Section 1.2.4, is used to test the null hypothesis $H_0 : \theta \leq 0$ against the alternative hypothesis $H_1 : \theta > 0$, with type I error probability $\alpha = 0.025$ and power $1 - \beta$ at $\theta = \delta$. Interim analyses are performed at fixed information levels $\mathcal{I}_1, \ldots, \mathcal{I}_{max}$, and the trial is stopped to reject H_0 if

$$Z_k = \hat{\theta}_k \sqrt{\mathcal{I}_k} \geq b_k \quad k = 1, \ldots, K.$$

Once the trial has started, the probability of falsely rejecting the null hypothesis $H_0 : \theta \leq 0$ can at any point be calculated, conditional on the data that has been observed so far. It is clear that for $\theta \leq 0$, this probability is maximised for $\theta = 0$. This conditional type I error probability will be referred to as $\mathrm{CRP}_{\theta=0}$ and plays a fundamental role for the CRP principle. Suppose that $Z_1 = z_1$ is observed at the first interim analysis. The conditional type I error can then be defined as

$$\mathrm{CRP}_{\theta=0} = \begin{cases} P_{\theta=0}\left\{\bigcup_{k=2}^{K}\left(\bigcap_{j=2}^{k-1} Z_j < b_j \text{ and } Z_k \geq b_k\right) | Z_1 = z_1\right\} \text{ if } Z_1 < b_1 \\ \\ 1 \text{ if } Z_1 \geq b_1. \end{cases} \tag{3.1}$$

The probability in (3.1) can now be calculated through a repeated evaluation of univariate integrals. Suitable methods for numerical integration are described in Jennison and Turnbull (2000, Chapter 19). The methods for numerical integration are based on the distribution theory for a sequence of estimates in a group sequential design, defined in equation (1.1) in Section 1.2.3. We also note that the calculation of (3.1) relies on the assumption that there is no uncertainty about which future information sequence will be observed.

The investigator is free to choose whether to continue with the original design, or re-design the remainder of the study as if it were a new trial with type I error probability equal to $\mathrm{CRP}_{\theta=0}$. We shall denote by $\tilde{\alpha}$ this type I error probability available for the re-design. Since the conditional type I error rate will be the same regardless of the investigator's choice, the overall type I error rate is also protected. This is in agreement with Jennison and Turnbull (2003), who show that any design which controls the overall type I probability and gives flexibility to choose whether to adapt or not, must also preserve the conditional type I error. A key feature of the CRP principle is that there

is almost complete flexibility for how to re-design the remainder of the trial. The only requirement is that the observations in the remainder of the trial must be independent from the observations that were used to calculate (3.1) and that the type I error of the new trial is at most $\tilde{\alpha}$. Denne (2001) also uses a conditional error function approach but requires that the adaptation takes place at the final interim analysis. The CRP principle is however more general. Although it was assumed in (3.1) that the re-design took place at one of the pre-planned interim analysis, Schäfer and Müller (2001) have shown that the re-design may occur at any point during a clinical trial. As described in Müller and Schäfer (2001), the principle can also be applied for two-sided tests. In this case, $\text{CRP}_{\theta=0}$ is calculated separately for the upper and lower boundaries, giving one value $\tilde{\alpha}_u$ for the upper boundary and another value $\tilde{\alpha}_l$ for the lower. Unless $\tilde{\alpha}_l = \tilde{\alpha}_u$, the two-sided test of the re-designed trial will thus be asymmetric, in the sense that a higher type I error can be spent for one of the boundaries than the other.

The CRP principle has gained popularity. It is implemented in the software package East-5 (2007) and has been applied in different areas. The method is used in the protocol of Martinez-Torres et al. (2008), while Scherag et al. (2009) report an application in genome-wide association studies. Rinke et al. (2009) provide results from a clinical trial in oncology, where re-design using the CRP principle was possible but not applied.

3.1.2 The CRP principle for unknown future information sequences

The CRP principle fits very naturally in a group sequential framework with fixed group sizes, but it would be very beneficial to be able to apply the method also when group sizes are unpredictable. Müller and Schäfer (2001) did not discuss how to calculate $\text{CRP}_{\theta=0}$ for unpredictable information sequences. Müller and Schäfer (2004) use an alpha spending function to derive the critical values for the original design, where it is planned that the interim analyses will be carried at certain information levels. When applying the CRP principle, the future group sizes are assumed to follow this original plan, but Müller and Schäfer (2004) do not discuss how to proceed if the observed group sizes would turn out to differ from this plan. In the user manual of the software package (East-5, 2007, p. 960), it is pointed out that the future course of the trial must be fully specified. No suggestions for how to handle deviations from this pre-specification are however given. Schäfer et al. (2006) point out that for a safe use of the CRP principle, the protocol must specify how to calculate $\text{CRP}_{\theta=0}$ at any time during the trial. As will be shown in this chapter, it may be challenging to construct such a rule in practice.

Several authors have discussed how to apply the CRP principle in situations where there is uncertainty about how to calculate $\text{CRP}_{\theta=0}$. In the context of a survival trial with uncertainty about the number of events at future interim analyses, Schäfer and Müller (2001) suggest specifying a rule for how to calculate $\text{CRP}_{\theta=0}$ in the protocol.

Posch et al. (2004) describe some issues when conditional rejection probabilities depend on nuisance parameters such as the unknown variance in a t-test, while Timmesfeld et al. (2006) give a very detailed account of the t-test situation. We consider the simple case of an error spending design with normally distributed data and known variance, as we are not aware of any publication that fully addresses how to calculate the conditional type I error probability in this situation.

Consider an error spending design with type I error spent in the usual way described in Section 1.2.5. Suppose that that at the first interim analysis, $Z_1 < b_1$, where b_1 is given by the error spending rule. To be able to apply the CRP principle, we need to calculate the conditional type I error. It is clear from (3.1) that to calculate the conditional type I error in a group sequential design, we need to know the information sequence that will actually be observed at future interim analyses. This sequence will in general not be known and may differ from the information sequence that was assumed when setting power, prior to the start of the trial. This does not cause problems in an error spending design without re-design, where $\mathcal{I}_k$ can be observed before deriving the critical value b_k for rejecting H_0.

In Section 3.2.2 we give an example of how substantially $\text{CRP}_{\theta=0}$ can depend on which future information sequence is used in the calculation. Thereafter, the impact on the overall type I error probability, when there is no pre-specification of how to calculate $\text{CRP}_{\theta=0}$, is assessed In Section 3.2.3. In Section 3.2.4, we investigate if the type I error can be controlled by using a pre-specified rule for calculation of $\text{CRP}_{\theta=0}$. In Section 3.3, adaptive designs with a pre-specified combination rule that controls the type I error are considered. We present efficiency comparisons between these adaptive designs and error spending designs, for the case when the future information sequence is unpredictable. Finally, conclusions about the possibilities for efficient and flexible designs with type I error control are presented in Section 3.4.

3.2 A numerical example

3.2.1 Framework

We consider the framework described in Section 3.1, where two treatments with normally distributed response and known variance are being compared in a clinical trial. Assume that the trial is monitored through the error spending method as described in Section 1.2.5, with interim analyses scheduled every six months until $\mathcal{I}_{max}$ has been reached. The type I error probability is spent according to

$$f(\mathcal{I}_k) = \begin{cases} 0.0001 & \text{for } k = 1 \\ \alpha \min\left((\mathcal{I}_k/\mathcal{I}_{max})^\rho, 1\right) & \text{for } k = 2, \ldots, K. \end{cases} \tag{3.2}$$

The wide boundary at the first interim analysis is plausible if results based on only a small amount of data may be unconvincing, unless they are overwhelmingly in favour of the alternative. On the other hand, a more reasonable chance of stopping for benefit at subsequent analyses, when more information has been accrued, can be achieved by setting $\rho = 1$.

Suppose that $\alpha = 0.025$ and that 90% power is required at $\theta = \delta = 0.1$. Based on these assumptions, we can calculate the information needed in a fixed sample trial as

$$\mathcal{I}_{fix} \;=\; \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta^2} = 1051.$$

For an error spending design with three equally spaced analyses and type I error spent according to (3.2) with $\rho = 1$, an inflation factor of $R = 1.08$ is required. For normally distributed data and two treatment groups of equal size with common known variance σ^2, the maximum information is converted to maximum sample size per treatment group according to

$$n_{max} = 2\sigma^2 \mathcal{I}_{max}.$$

We thus obtain that the maximum information $\mathcal{I}_{max} = 1135$ and that for $\sigma^2 = 1$, $n_{max} = 2270$ patients per treatment group are required.

3.2.2 Conditional type I error for the re-designed trial

Suppose that it turns out that at the time of the first interim analysis, the investigators find it desirable to perform a re-design, by applying the CRP principle. To calculate $\tilde{\alpha} = \text{CRP}_{\theta=0}$, assumptions about the future information sequence are however needed. Suppose that recruitment has been slower than expected and that the information observed is $\mathcal{I}_1 = \mathcal{I}_{max}/5$, rather than the planned $\mathcal{I}_{max}/3$. With equally spaced analyses in time, it seems very natural to assume that the observed accrual rate will continue until $\mathcal{I}_{max}$ has been reached and use

$$\mathcal{I}^{p_1} \;=\; \mathcal{I}_{max} \;\times\; (0.4, 0.6, 0.8, 1.0).$$

Another reasonable choice would be to suppose that future information sequences will follow the plan specified prior to the start of the trial, i.e. to use

$$\mathcal{I}^{p_2} \;=\; \mathcal{I}_{max} \;\times\; (0.67, 1.0).$$

Using $\mathcal{I}^{p_2}$ bears similarities with the rule used by Schäfer and Müller (2001), who stipulate that $\text{CRP}_{\theta=0}$ in a survival trial will be calculated according to the planned number of events at future interim analyses. As a third possibility, we consider

$$\mathcal{I}^{p_3} \;=\; \mathcal{I}_{max} \;\times\; (1.0),$$

Information sequence	$\tilde{\alpha} = \text{CRP}_{\theta=0}$
$\mathcal{I}^{p_1} = \mathcal{I}_{max} \times (0.40, 0.60, 0.80, 1.0)$	0.64
$\mathcal{I}^{p_2} = \mathcal{I}_{max} \times (0.67, 1.0)$	0.43
$\mathcal{I}^{p_3} = \mathcal{I}_{max} \times (1.0)$	0.32

Table 3.1: Conditional type I error for different future information sequences, when $\mathcal{I}_1 = 0.2\mathcal{I}_{max}$ and $Z_1 = 3.47$.

as it has been used by other authors like Cui et al. (1999) as a simple default rule for calculating conditional power.

There could be different reasons for wishing to apply the CRP principle in this situation. Suppose that the treatment effect is estimated to be $\hat{\theta} = 0.23$, i.e. more than double what was assumed in setting power at the design stage. As

$$Z_1 = \hat{\theta}\sqrt{(n_1/2\sigma^2)} = 3.47,$$

the boundary of $b_1 = 3.719$ is almost crossed. The investigators may wish to design a new trial with more frequent interim analyses. Indeed, Müller and Schäfer (2001) point to the possibility of changing the number and timing of interim analyses as one of the major benefits of their method. Another reason for re-design could be the wish to rescue an under-powered study, if the treatment effect turns out to be worse than expected. External information that has emerged about a competitor drug could be yet another reason to re-design, even though the treatment effect is largely as expected.

Table 3.1 lists $\text{CRP}_{\theta=0}$, calculated for $\mathcal{I}^{p_1}$, $\mathcal{I}^{p_2}$ and $\mathcal{I}^{p_3}$, when $\mathcal{I}_1 = 0.2\mathcal{I}_{max}$ and $Z_1 = 3.47$. The results in Table 3.1 vary substantially depending on which future information sequence is used. The conditional type I error for information sequence $\mathcal{I}^{p_1}$ is, for example, about two times that calculated for $\mathcal{I}^{p_3}$. Investigators have to make a choice about which future information sequence to use. Choosing an option which gives higher power is attractive, so one possibility is to choose the information sequence that gives the highest $\text{CRP}_{\theta=0}$, i.e. $\mathcal{I}^{p_1}$ in this case. Hence, there seems to be scope for increasing power by choosing a "suitable" future information sequence when calculating $\text{CRP}_{\theta=0}$. This possibility of increasing power also raises the question of whether the overall type I error probability can be inflated.

3.2.3 Impact on overall type I error

Suppose it is decided to re-design a trial at the first interim analysis by applying the CRP principle, but there is uncertainty about which future information sequence will be observed if the study continues as originally planned. Consider future information

sequences $\mathcal{I}^{p_1}, \ldots, \mathcal{I}^{p_N}$ that seem plausible; here each $\mathcal{I}^{p_j}$ is a sequence of values for $\mathcal{I}_2$ to $\mathcal{I}_{\tilde{K}}$, where $\tilde{K}$ is the number of analysis needed to reach $\mathcal{I}_{max}$. It may well be tempting to increase the probability of a positive study, by choosing the information sequence, referred to as $\mathcal{I}^{p_{max}}(z_1)$, that among different future information sequences $\mathcal{I}^{p_j}$ maximises

$$\mathrm{CRP}_{\theta=0}\left(\mathcal{I}^{p_j}, Z_1 = z_1\right), \tag{3.3}$$

the conditional type I error calculated for $Z_1 = z_1$. If $\phi(z)$ is the standard normal probability density function, the overall type I error for the procedure is given by

$$\alpha_i\left(\mathcal{I}^{p_{max}}(z_1)\right) \;=\; \int_{-\infty}^{\infty} dz_1 \phi(z_1) \mathrm{CRP}_{\theta=0}\left(\mathcal{I}^{p_{max}}(z_1), z_1\right). \tag{3.4}$$

The integral in (3.4), as well as the other integrals below, can be calculated through numeric integration using the methodology described by Jennison and Turnbull (2000, Chapter 19). We note that

$$\int_{-\infty}^{\infty} dz_1 \phi(z_1) \mathrm{CRP}_{\theta=0}\left(\mathcal{I}^{p_j}, z_1\right) = \alpha, \tag{3.5}$$

for $j = 1, \ldots, N$, while

$$\alpha_i\left(\mathcal{I}^{p_{max}}(z_1)\right) = \int_{-\infty}^{\infty} dz_1 \phi(z_1) \max_{j=1,\ldots,N} \mathrm{CRP}_{\theta=0}\left(\mathcal{I}^{p_j}, z_1\right) \geq \alpha. \tag{3.6}$$

Since the integrand in (3.6) is continuous, the integral will equal α only if

$$\mathrm{CRP}_{\theta=0}(\mathcal{I}^{p_{max}}(z_1), z_1) = \mathrm{CRP}_{\theta=0}(\mathcal{I}^{p_j}, z_1) \quad j = 1, \ldots, N \tag{3.7}$$

for all z_1. The condition requires that $\mathrm{CRP}_{\theta=0}(\mathcal{I}^{p_j}, z_1)$ are equal for all j at every z_1. If $\mathcal{I}^{p_{max}}(z_1)$ is consistently used to calculate the type I error probability $\tilde{\alpha}$, to be used for the re-design of the remainder of the trial according to the CRP principle, then clearly the overall type I error will no longer be protected.

The conditional type I error probabilities for our three future information sequences $\mathcal{I}^{p_1}$, $\mathcal{I}^{p_2}$ and $\mathcal{I}^{p_3}$, calculated after having observed $Z_1 = z_1$ and $\mathcal{I}_1 = 0.2\mathcal{I}_{max}$, are displayed in Figure 3-1. Figure 3-1 illustrates that the maximum of the three curves is noticeably higher than each individual curve, by a small amount over a region with high $\phi(z_1)$ or a large amount over a region with low $\phi(z_1)$. If the results at the first interim analysis are positive, like in Table 3.1, a higher conditional type I error is

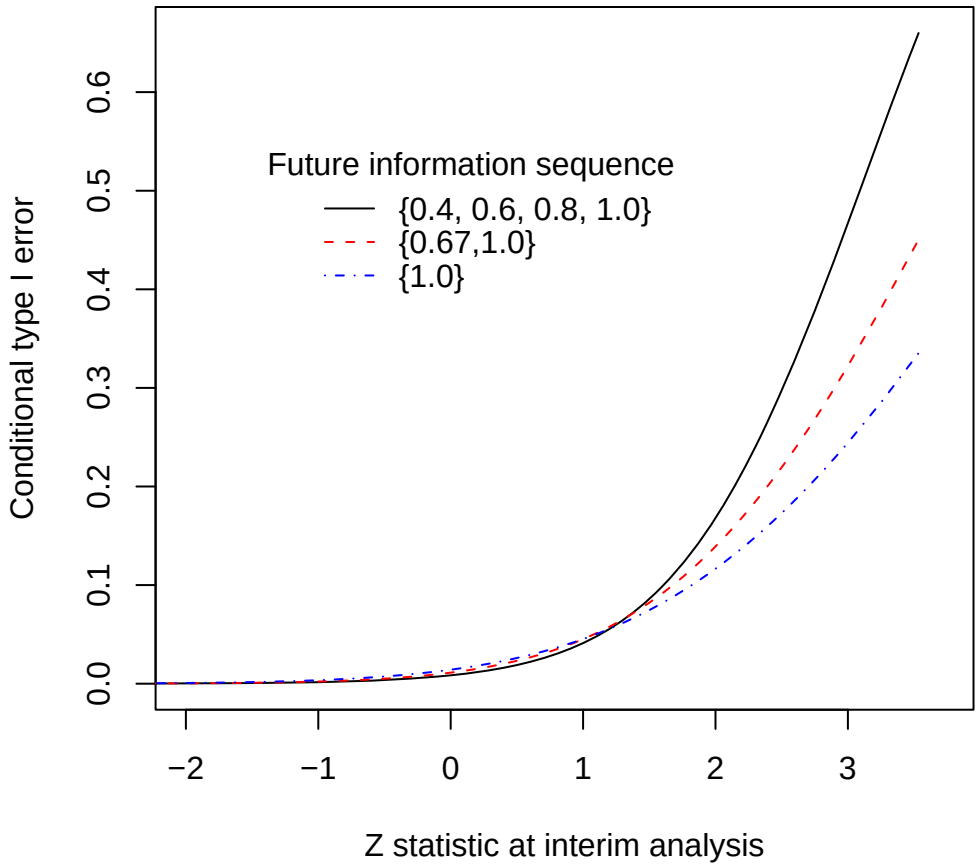

Figure 3-1: Conditional type I error for different future information sequences, after having observed $Z_1 = z_1$ at $\mathcal{I}_1 = 0.2\mathcal{I}_{max}$.

achieved for $\mathcal{I}^{p_1}$ than for $\mathcal{I}^{p_2}$ and $\mathcal{I}^{p_3}$. If results go in the other direction, it is $\mathcal{I}^{p_3}$ that gives the highest conditional type I error probability.

By considering the information sequences $\mathcal{I}^{p_1}$, $\mathcal{I}^{p_2}$, and $\mathcal{I}^{p_3}$, we can now calculate the overall type I error according to (3.6), where $\text{CRP}_{\theta=0}$ is always calculated using the information sequence that maximises the conditional type I error. The calculation can be performed for different α spending functions and for maximisation over different sets of information sequences. If for example $\mathcal{I}^{p_2}$ is thought to be hard to justify, the type I error can still be inflated by considering $\mathcal{I}^{p_1}$ and $\mathcal{I}^{p_3}$. As we would expect, the type I error turns out to be smaller if only two of the three information sequences are considered. The type I error is however larger than 0.025 provided that the maximisation is carried out over more than one information sequence. These results

Information sequences	$f(\mathcal{I}_1) = 0.0001$			$f(\mathcal{I}_1) = \alpha \min((\mathcal{I}_1/\mathcal{I}_{max})^\rho, 1)$		
maximised over	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 1$	$\rho = 2$	$\rho = 3$
$\mathcal{I}^{p_1}, \mathcal{I}^{p_2}, \mathcal{I}^{p_3}$	0.0289	0.0270	0.0261	0.0273	0.0267	0.0261
$\mathcal{I}^{p_1}, \mathcal{I}^{p_2}$	0.0271	0.0259	0.0254	0.0261	0.0257	0.0254
$\mathcal{I}^{p_1}, \mathcal{I}^{p_3}$	0.0288	0.0269	0.0261	0.0273	0.0267	0.0261
$\mathcal{I}^{p_2}, \mathcal{I}^{p_3}$	0.0268	0.0261	0.0257	0.0262	0.0260	0.0256
$\mathcal{I}^{p_1}$	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
$\mathcal{I}^{p_2}$	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250
$\mathcal{I}^{p_3}$	0.0250	0.0250	0.0250	0.0250	0.0250	0.0250

Table 3.2: Type I error depending on ρ family alpha spending function, type I error spent at first interim analysis and the information sequences considered to calculate $\text{CRP}_{\theta=0}$.

are summarised in Table 3.2, for ρ family error spending functions with $\rho = 1$, 2 and 3. Table 3.2 shows that when the type I error is spent according to (3.2) with $\rho = 1$, the type I error is in the range of $0.027 - 0.029$, depending on the future information sequences that are considered. We also present results for a slightly modified spending function, when type I error spent at the first interim analysis is $\alpha \min((\mathcal{I}_1/\mathcal{I}_{max})^\rho, 1)$, rather than 0.0001. The columns to the right in Table 3.2 show results for this case.

Further results are displayed in Figures 3-2 and 3-3. In Figure 3-2, we see that the type I error inflation is monotonically decreasing in ρ. This can be understood by noting that if ρ is high, a large proportion of the type I error is spent at the final analysis. The information sequences that we have considered all assume that the planned $\mathcal{I}_{max}$ is reached at the final analysis, so for high ρ most of the type I error is spent there. Hence, the conditional type I error for the different information sequences will be more similar, than if ρ is low and a larger proportion of the type I error is spent before the final analysis. This in turn limits the potential impact on the overall type I error that can be achieved by calculating $\text{CRP}_{\theta=0}$ for $\mathcal{I}^{p_{max}}(z_1)$, the information sequence that maximises the conditional type I error.

The situation is different in Figure 3-3, where the type I error spent at the first interim analysis depends on ρ, and increases with decreasing ρ. The type I error spent at the first interim analysis cannot contribute to differences in $\text{CRP}_{\theta=0}$, which makes the potential impact on the type I error by using $\mathcal{I}^{p_{max}}(z_1)$ to calculate $\text{CRP}_{\theta=0}$ smaller. So when $\alpha \min((\mathcal{I}_1/\mathcal{I}_{max})^\rho, 1)$ is spent at the first interim analysis, low values of ρ contribute to decreasing the overall type I error a little compared to the results in Figure 3-2, where the type I error spent at the first interim analysis equals 0.0001.

As would be expected from (3.5), the type I error is controlled at $\alpha = 0.025$ in both Figure 3-2 and Figure 3-3, provided that $\text{CRP}_{\theta=0}$ is always calculated for the same

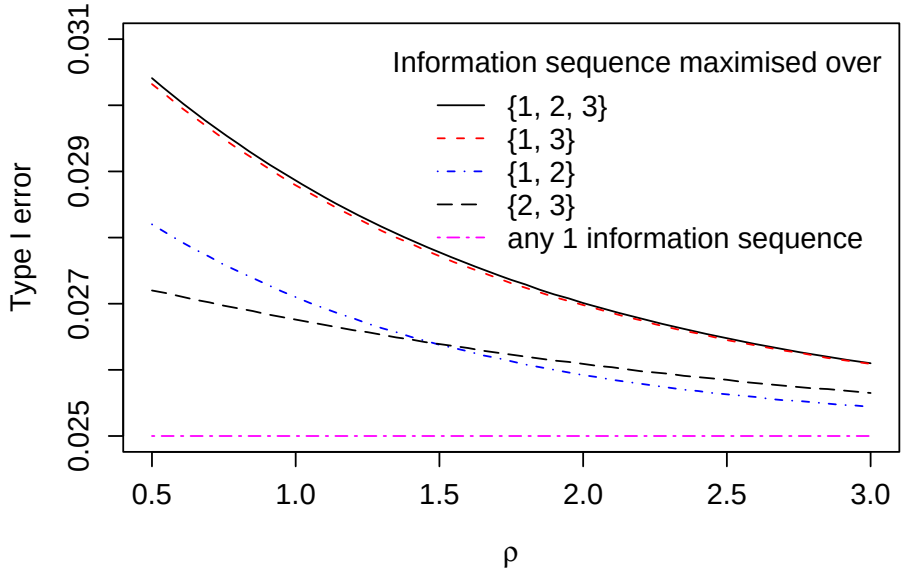

Figure 3-2: Overall type I error depending on ρ family alpha spending function and the information sequences considered to calculate $\text{CRP}_{\theta=0}$. The type I error spent at the first interim analysis equals 0.0001.

future information sequence. This suggests that always using the same sequence for the calculation might be a way of avoiding type I error inflation. We shall investigate this approach in the next section.

3.2.4 Calculating the conditional type I error according to a pre-specified rule

Suppose that it is pre-specified that the conditional type I error will always be calculated according to the plan in the protocol. Such an approach could be applied for our example with three equally spaced analyses. For a re-design at analysis 1 it may for example be specified that $\text{CRP}_{\theta=0}$ will be calculated assuming that the second interim analysis will occur at $2\mathcal{I}_{max}/3$, with a final analysis at $\mathcal{I}_{max}$. A similar approach is suggested by Schäfer and Müller (2001). Consider the possibility that more information than expected, $\mathcal{I}_{max}/2$ instead of $\mathcal{I}_{max}/3$, is observed at the first interim analysis. As before, the investigator needs to decide whether to use the CRP principle and re-design the remainder of the trial with type I error $\tilde{\alpha}$. The possible decisions are

$$D_1 \quad : \quad \text{Continue without re-design, or}$$
$$D_2 \quad : \quad \text{Re-design.}$$

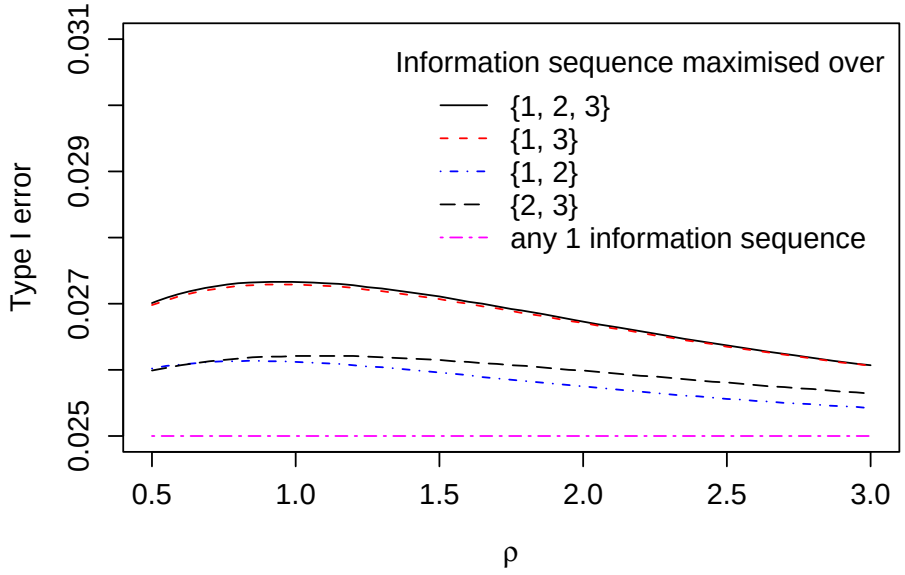

Figure 3-3: Overall type I error depending on ρ family alpha spending function and the information sequences considered to calculate $\text{CRP}_{\theta=0}$. The type I error spent at the first interim analysis equals $\alpha \min((\mathcal{I}_1/\mathcal{I}_{max})^{\rho}, 1)$.

If the study continues without re-design, a very reasonable assumption is that $\mathcal{I}_{max}$ will be reached after another six months, i.e. at the next interim analysis. It is clear that choosing D_1 will give a different $\text{CRP}_{\theta=0}$ than D_2. Suppose that it is decided to make the decision that gives the highest conditional type I error probability. Basing the decision D on the conditional type I error probabilities for D_1 and D_2 leads to the decision rule

$$D = \begin{cases} D_1 & \text{if } \text{CRP}_{\theta=0}(D_1) \geq \text{CRP}_{\theta=0}(D_2) \\ D_2 & \text{if } \text{CRP}_{\theta=0}(D_2) > \text{CRP}_{\theta=0}(D_1). \end{cases} \tag{3.8}$$

We find that if the decision rule in (3.8) is used to decide whether to re-design, the type I error can be inflated from 0.025 to 0.0288, i.e. an inflation of similar size as the results in Table 3.2.3. An even more extreme situation occurs if $\mathcal{I}_1 = 0.6\mathcal{I}_{max}$, in which case application of (3.8) can inflate the type I error from 0.025 to 0.0302. It may be argued that it is impossible for the investigator to know exactly which future information sequence will be observed if D_1 is chosen. In many situations, including the present example, it should however be fairly obvious that the future information sequence, in the case of no re-design, will differ substantially from what was specified in the protocol. Furthermore, the CRP principle may be applied repeatedly, at any

time during the trial. Moreover, we have seen cases with even higher inflation, if we do not require that $\mathcal{I}_{max}$ is reached exactly at the final analysis for all the information sequences that are considered. Hence, type I error inflation is possible even when a rule in the protocol stipulates how to calculate $\text{CRP}_{\theta=0}$. The increase in type I error is obtained by choosing between the decisions D_1 and D_2, using the decision rule defined in (3.8). In summary, we would envisage that most of the inflation in type I error probability seen in this example could be obtained in practice.

3.3 A method that controls type I error exactly

3.3.1 Weighted inverse normal method

It is clear that type I error control is an issue whenever there is a possibility to apply the CRP principle, and $\text{CRP}_{\theta=0}$ is not uniquely defined. One solution is to use a pre-specified combination rule of the type proposed by Bauer and Köhne (1994), which is described in Section 1.3.2. We shall illustrate how to apply the CRP principle with another pre-specified combination rule, the weighted inverse normal method of Lehmacher and Wassmer (1999), which was briefly introduced in Section 1.3.3.

The method of Lehmacher and Wassmer (1999) can be illustrated by considering a group sequential trial with K groups of observations. Suppose that data are normally distributed and let $\tilde{Z}_k$ denote the standardised statistic based on data from stage k of the trial. The sequence of test statistics

$$Z_M = \frac{\sum_{k=1}^{M} w_k \tilde{Z}_k}{(\sum_{k=1}^{M} w_k^2)^{1/2}}, \quad M = 1, \ldots, K, \tag{3.9}$$

can then be defined. The decision whether to stop or continue is based on these test statistics, and Lehmacher and Wassmer (1999) show that data-dependent sample size modification is possible without inflating the type I error rate. The weights w_k in (3.9) may be chosen to reflect the pre-planned groups and need not be equal, but Lehmacher and Wassmer (1999) require fixing the weights prior to the start of the trial. Equally informative observations may thus be given different weights, if the observed group sizes differ from what was planned at the outset. Marginally, each $\tilde{Z}_k$ follows a $N(\theta\sqrt{\mathcal{I}_k - \mathcal{I}_{k-1}}, 1)$ distribution, where $\mathcal{I}_k$ is Fisher's information for θ at analysis k and $\mathcal{I}_0 = 0$. The sequence of test statistics $Z_1, \ldots, Z_K$ in (3.9) then has the following

properites:

$$
\begin{aligned}
&(Z_1, \ldots, Z_K) \text{ is multivariate normal,} \\
&E(Z_M) = \theta \sum_{k=1}^{M} \frac{\sqrt{\mathcal{I}_k - \mathcal{I}_{k-1}} w_k}{(\sum_{k=1}^{M} w_k^2)^{1/2}}, \quad k = 1, \ldots, K, \text{ and} \\
&\mathrm{Cov}(Z_{M_1}, Z_{M_2}) = \sqrt{\frac{\sum_{k=1}^{M_1} w_k^2}{\sum_{k=1}^{M_2} w_k^2}}, \quad 1 \leq M_1 \leq M_2 \leq K.
\end{aligned}
\tag{3.10}
$$

There are similarities with the standard joint canonical distribution in (1.1), but the correlation structure is now decided by the pre-specified weights in (3.9) rather than by the observed group sizes. In order to preserve the type I error rate, critical values for rejecting the null hypothesis can be derived in the same way as for non-adaptive group sequential designs with fixed information sequences, using the fact that the correlation structure is known from (3.10) and the pre-specified weights. This correlation structure can also be used to calculate the conditional type I error probability. In addition, the critical values do not change because of the observed group sizes. Hence, the conditional type I error is well defined and we can apply the CRP principle, without facing the problems with how to calculate $\mathrm{CRP}_{\theta=0}$ encountered in the previous sections.

Let us now return to the 3-group error spending design described in Section 3.2.2, where we test the null hypothesis $\theta \leq 0$, with type I error $\alpha = 0.025$ and 90% power is required at $\theta = \delta = 0.1$. The type I error is spent according to (3.2) and three equally spaced analyses are planned, at

$$\mathcal{I}_k = \frac{k}{K} \times \mathcal{I}_{max}, \quad k = 1, \ldots, K$$

for $\mathcal{I}_{max} = 1135$. Suppose that prior to the start of the trial, the weights in (3.9) have been specified as $w_1 = w_2 = w_3 = \frac{1}{3}$, to reflect the pre-planned group sizes. The conditional type I error is then unambiguously defined and equals 0.71, when calculated according to (3.1) for $\mathcal{I}_1 = 0.2\mathcal{I}_{max}$ and $Z_1 = 3.47$. The higher conditional type I error compared to the numbers in Table 3.2.2 is to be expected, as more weight is given to the first stage data than for an error spending design that adapts to the observed information of $\mathcal{I}_1 = 0.2\mathcal{I}_{max}$.

3.3.2 Efficiency comparison with error spending design

Lehmacher and Wassmer (1999) present efficiency comparisons between their method, applied with 5 groups of observations, and 5-group Pocock (1977) tests. The comparisons focus on the situation when there is no design adaptation and the initially planned group sizes have been perturbed. In this situation Lehmacher and Wassmer (1999) find that there is some efficiency loss associated with their method, compared

to the Pocock test. The critical values of the Pocock tests are in these comparisons derived assuming that the observed group sizes that will later be observed are known at the outset. This is necessary to obtain Pocock tests with exact control of the type I error rate, as the Pocock test cannot really deal with perturbations of the group sizes that are unknown when planning the design. The comparison may thus be regarded as unfair to the method of Lehmacher and Wassmer (1999), as in practice the group sizes that will later be observed would not be known when planning the Pocock test.

We shall instead make a comparison with error spending designs, as these just like the method of Lehmacher and Wassmer (1999) control type I error exactly also for unpredictable information sequences. Unpredictable information sequences can however be an issue also for error spending designs, as perturbation of the initially planned group sizes does have a small impact on power and expected sample size. Hence, it is of interest to assess how the efficiencies of the method of Lehmacher and Wassmer (1999) and the error spending approach as described by Jennison and Turnbull (2000, Chapter 7) compare, when having to deal with unpredictable group sizes.

Let us again consider the numerical example in Section 3.2, with $\mathcal{I}_{fix} = 1051$. We now consider an error spending design from the ρ family, planned for $K = 5$ equally spaced analyses. For the method of Lehmacher and Wassmer (1999), the pre-specified weights are set to $1/5$ for all five stages. This choice reflects the initial plan of five equally spaced analyses. For $\rho = 1.5$, we obtain $R = 1.086$ and $\mathcal{I}_{max} = R \times \mathcal{I}_{fix} = 1141$. The boundary $b_1, \ldots, b_5$ can be derived from the error spending function and the planned information sequence, and is the same for both the error spending design and the Lehmacher and Wassmer design. Our efficiency comparison will focus on the situation where no design changes take place, but there is some perturbation of the initially planned information sequence. The observed information sequences considered are of the form

$$\mathcal{I}_k = \left(\frac{k}{K}\right)^s \times \mathcal{I}_{max}, \quad k = 1, \ldots, K,$$

where $s > 0$. We note that for $s = 1$, the observed group sizes coincide with the pre-planned weights and the multivariate normal distribution of the sequence of statistics in (3.10) coincides with that in (1.1) obtained for classical group sequential tests. For $s = 0.6$, we obtain the information sequence

$$\mathcal{I}_{max} \times (0.38, 0.58, 0.74, 0.87, 1.0),$$

while for $s = 1.6$, we obtain

$$\mathcal{I}_{max} \times (0.08, 0.23, 0.44, 0.70, 1.0).$$

Whenever the parameter $s \neq 1$, the boundary of the error spending design will be

adjusted, while the boundary of the Lehmacher and Wassmer design will remain the same. The statistics $\tilde{Z}$ in the Lehmacher and Wassmer design are weighted according to the pre-specified weights when calculating the pooled statistic Z_M in (3.9), even though the group sizes are not the same for all k. The two methods will have changes in both expected sample size and power, so focusing on just one of these two properties will not give the complete picture. To enable a fair assessment of the methods, the efficiency index proposed by Jennison and Turnbull (2006a) is used. Suppose that at treatment effect θ, an error spending design with type I error rate α has power $1 - b_{\mathrm{ES}}(\theta)$ and expected information $\mathrm{E}_{\mathrm{ES},\theta}(\mathcal{I})$. In a similar way, a Lehmacher and Wassmer design is assumed to have power $1 - b_{\mathrm{LW}}(\theta)$ and expected information $\mathrm{E}_{\mathrm{LW},\theta}(\mathcal{I})$ at θ. We can then define efficiency indexes $\mathrm{EI}_{\mathrm{ES}}(\theta)$ and $\mathrm{EI}_{\mathrm{LW}}(\theta)$ according to

$$\mathrm{EI}_{\mathrm{ES}}(\theta) = \frac{(z_{1-\alpha} + z_{b_{\mathrm{ES}}(\theta)})^2}{\theta^2} \frac{1}{\mathrm{E}_{\mathrm{ES},\theta}(\mathcal{I})} \tag{3.11}$$

$$\mathrm{EI}_{\mathrm{LW}}(\theta) = \frac{(z_{1-\alpha} + z_{b_{\mathrm{LW}}(\theta)})^2}{\theta^2} \frac{1}{\mathrm{E}_{\mathrm{LW},\theta}(\mathcal{I})} \tag{3.12}$$

and an efficiency ratio according to

$$\mathrm{ER}_{\mathrm{LW/ES}}(\theta) = 100 \times \frac{\mathrm{EI}_{\mathrm{LW}}(\theta)}{\mathrm{EI}_{\mathrm{ES}}(\theta)} = 100 \times \frac{(z_{1-\alpha} + z_{b_{\mathrm{LW}}(\theta)})^2}{(z_{1-\alpha} + z_{b_{\mathrm{ES}}(\theta)})^2} \frac{\mathrm{E}_{\mathrm{ES},\theta}(\mathcal{I})}{\mathrm{E}_{\mathrm{LW},\theta}(\mathcal{I})}. \tag{3.13}$$

Figure 3-4 shows the power and expected sample size at $\theta = \delta$, as well as the efficiency ratio, defined in (3.13) and also calculated at $\theta = \delta$. As a benchmark, we also include horizontal lines for the power, expected sample size and efficiency ratio, for the case when there is no perturbation to the initially planned information sequence. The power of the error spending design is close to the target of 0.9, which is in agreement with the findings of Jennison and Turnbull (2000, Chapter 7). For $0.3 \leq s \leq 1.3$, the power of the method of Lehmacher and Wassmer (1999) is also close to its target value, while the power loss is more pronounced for extreme values of s. The expected sample size is lower for the error spending design for some values of s and for the method of Lehmacher and Wassmer (1999) for others. Hence, considering the efficiency ratio should provide a better assessment of the relative efficiency of the methods than considering power or expected sample size in isolation. The right panel of Figure 3-4 shows that $95 \leq \mathrm{ER}_{\mathrm{LW/ES}}(\theta = \delta) \leq 100$, for $0.6 \leq s \leq 1.6$. If the efficiency of the designs is judged by $\mathrm{ER}_{\mathrm{LW/ES}}(\theta = \delta)$, there is thus some efficiency loss associated with the method of Lehmacher and Wassmer. On the other hand, if re-design is desired, the pre-specified combination rule gives the increased flexibility of being able to apply the CRP principle, with a well-defined rule for how to calculate $\mathrm{CRP}_{\theta=0}$. The size of the efficiency loss depends on to what degree the parameter s differs from one. In practice, the observed group sizes will not necessarily follow a neat pattern like the

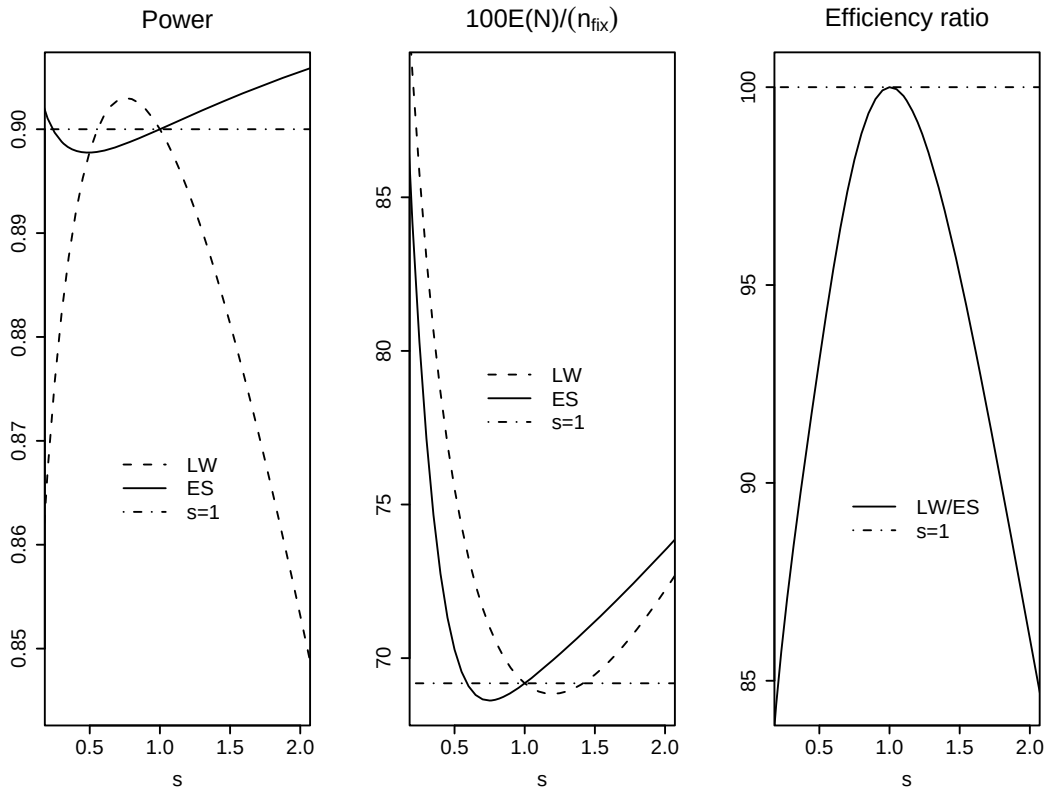

Figure 3-4: Power, $100E(N)/n_{fix}$ and efficiency ratio at $\theta = \delta$, for a 5-group, Lehmacher and Wassmer (dashed line) design and a 5-group, $\rho = 1.5$ error spending design (solid line). Also shown are horizontal dot-dashed lines with efficiency ratio, expected sample size and power at $\theta = \delta$, for the case of no perturbations (i.e. s=1).

ones considered here. If the perturbations are of roughly the same size as for $s = 0.9$ or $s = 1.1$, the efficiency loss of the method of the method of Lehmacher and Wassmer (1999) compared to the error spending design is quite small. When choosing between using an error spending design and the method of Lehmacher and Wassmer (1999), there appears to be a trade-off between efficiency and flexibility. For values of s close to one, it may well be that the higher flexibility of the method of Lehmacher and Wassmer (1999) compensates for the small efficiency loss that is incurred.

3.4 Discussion

In many situations it can be difficult to predict the amount of information that will be available at future interim analyses. Maximum information error spending designs provide a very flexible method for coping with unpredictable information sequences without inflation of the overall type I error probability. Furthermore, error spending designs have been shown to provide efficient designs in many situations, see for example

Barber and Jennison (2002) for comparisons against optimal non-adaptive group-sequential designs and Jennison and Turnbull (2006b) for comparisons against optimal adaptive group sequential designs.

According to Müller and Schäfer (2001, p. 890), "any group sequential design can be given the full flexibility of an adaptive design". We have however shown that error spending designs cannot be given the full flexibility of adaptive designs, without losing some of their most important properties. This is a concern, as error spending designs are widely used, and it would be very appealing to be able to combine the two approaches. The numerical examples in Section 3.2 clearly show that the overall type I error probability is not protected, unless the conditional type I error probability is uniquely defined in all situations when a re-design can occur. In particular, the example in Section 3.2.4 shows that there are even situations where the approach of pre-specifying how to calculate $\text{CRP}_{\theta=0}$ does not guarantee type I error control.

In the numerical examples, it was shown that the type I error can be inflated from 0.025 to about 0.029, while an admittedly rather extreme example produced an inflation from 0.025 to just over 0.03. It should also be mentioned that while only one re-design point has been considered in our examples, recursive use of the CRP principle may lead to additional inflation. It is well known that error spending tests protect the type I error rate exactly, provided that the future information sequence is independent of the observed treatment effect. This important property is however lost if the CRP principle is applied in situations when $\text{CRP}_{\theta=0}$ is ambiguously defined.

In Section 3.3 it is shown that the type I error is controlled when a pre-specified combination rule for multi-stage defines how to calculate $\text{CRP}_{\theta=0}$. In contrast to the methods that require pre-specified weights, the error spending tests adjusts to the observed group sizes and all observations are weighted equally, which in Figure 3-4 is shown to be more efficient. Applying the CRP principle in a design with a pre-specified combination rule, like the weighted inverse normal method of Lehmacher and Wassmer (1999), must be seen as an alternative to error spending tests, rather than as a method that enjoys all the benefits of error spending tests plus additional advantages.

As discussed in Section 3.2.4, inflation of the type I error rate is possible if the investigators may choose between two conditional type I error probabilities that are not equal at all times during the trial. Suppose that an independent body, with full access to all trial data but without access to individuals involved in running the trial, is given the task of estimating the future information sequence that is most likely to occur. This independent body could for example be a regulatory authority. If the investigator were interested in designing a new trial according to the CRP principle, $\text{CRP}_{\theta=0}$ would simply be calculated by the independent body. Such an approach would limit the possibility of inflating the overall type I error probability by using the decision rule defined in (3.8). A more conservative procedure would be to consider a class of

future information sequences that are plausible and use the information sequence that gives the lowest $\text{CRP}_{\theta=0}$. A drawback with such an approach is that some conservatism in type I error probability would be inevitable as a result.

If a decision is made to use the CRP principle in a clinical trial, a simple approach that protects the type I error rate is given by pre-specified combination rules like that of Lehmacher and Wassmer (1999). Such an approach can lead to some efficiency loss for unpredictable information sequences, while the type I control is controlled exactly. When information sequences are unpredictable, the type I error inflation that can occur when there is no pre-specified combination rule that uniquely defines $\text{CRP}_{\theta=0}$, is potentially more serious than the efficiency loss of the weighted inverse normal method. The recursive combination tests of Brannath et al. (2002) provide another alternative that ensures control of the type I error, even though repeated use of the method may be logistically challenging. The possibility of giving an independent body the task of estimating future information sequences also deserves further consideration. Most importantly, the pros and cons should be compared to standard error spending designs before a decision is made to use the CRP principle in a clinical trial.

In summary, the CRP principle provides a very interesting and flexible tool for clinical trial design. It is however crucial that the potential issues highlighted in this chapter are taken into account at the design stage and dealt with in the study protocol. The increased flexibility that the method provides comes at a price. This price is paid either in terms of credibility of results when there may have been inflation of the type I error rate, or in terms of some efficiency loss.

CHAPTER 4

Group sequential designs with non-binding futility boundaries

4.1 Introduction

In the previous chapter we saw that type I error control can be a concern when applying novel adaptive methods. We will now consider a situation where type I error control can be an issue also for classical group sequential designs. The problem that we will consider arises for one-sided group sequential tests with futility stopping, so it is not applicable for the group sequential designs without futility boundary in Chapter 3. We will start by giving a brief introduction to one-sided group sequential designs with futility boundaries, and then move on to discuss issues about type I error control if there is a concern that the futility boundary will not always be applied. As a solution, we will consider making the futility boundaries non-binding. With a non-binding futility boundary, the upper boundary is set to protect the type I error rate, even if the study continues after the futility boundary has been crossed and the null hypothesis is subsequently rejected. Later in the chapter we will present a new method for deriving optimal group sequential designs with non-binding futility boundaries, and compare these optimal designs to other designs in the same class.

Consider a clinical trial, comparing an experimental drug to placebo or an active control, that is monitored using group sequential methodology. It is then typically possible to stop early for benefit, when at an interim stage the estimated treatment difference is over-whelmingly in favour of the experimental drug. It may however also be preferable to stop early if, based on interim data, the trial is unlikely to deliver the benefits hoped for at the outset. Early stopping in the latter situation is often referred to as stopping for futility and it is clear that many benefits can be achieved by such a decision. As discussed in Chapter 1, there are both ethical and economical reasons to stop. Patients do not have to be exposed to a drug that is unlikely to be effective, and may have unexpected side effects. Moreover, the trial sponsor can usually save some

of the resources that would have been needed to complete the trial. Focus can instead be shifted to study another promising treatment that in light of the interim trial data is more likely to bring benefit to patients.

When the primary outcome of a clinical trial is monitored within a group sequential framework, it is generally accepted that certain adjustments have to be made to account for the multiple tests that are performed. In particular, trials that form the basis for a new drug application for regulatory approval must meet a certain overall type I error probability, typically $\alpha = 0.025$ for a one-sided test. While it is generally straightforward to use numerical integration techniques to calculate the probability of falsely rejecting the null hypothesis given a certain stopping rule, the actual conduct of clinical trials makes things more complicated. We shall first describe the general framework of a one-sided group sequential test in the clinical trial setting, and thereafter point to some issues regarding how to calculate the overall type I error probability of such trials.

Suppose that in a clinical trial comparing two treatments, the observations X_{Aj} and X_{Bj}, $j = 1, 2, \ldots,$ on treatments A and B respectively, are independent and normally distributed with $X_{Aj} \sim N(\mu_A, \sigma^2)$ and $X_{Bj} \sim N(\mu_B, \sigma^2)$, where the common variance σ^2 is known. We wish to make inference about the difference in means $\theta = \mu_B - \mu_A$, by performing a one-sided K - group sequential test of $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. Denote by n_k the cumulative per-group sample size at interim analysis k, for $k = 1, \ldots, K$. Given information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$, where $\mathcal{I}_k = n_k/(2\sigma^2)$, we can find a stopping boundary so that the test has type I error probability α and power $1 - \beta$ at $\theta = \delta$.

Let $\hat{\theta}_k$ be the maximum likelihood estimate of θ at analysis k. At each interim analysis, we calculate the usual standardised statistic

$$Z_k = \hat{\theta}_k \sqrt{\mathcal{I}_k}$$

and stop early to reject H_0 if $Z_k \geq b_k$ and to accept H_0 if $Z_k \leq a_k$. Termination is ensured by setting $a_K = b_K$ at analysis K. The type I error and the power at $\theta = \delta$, respectively, are defined according to

$$\begin{aligned} P_{\theta=0}(\text{Reject } H_0) &= P_{\theta=0}(\text{Cross upper boundary before lower boundary}) \\ P_{\theta=\delta}(\text{Reject } H_0) &= P_{\theta=\delta}(\text{Cross upper boundary before lower boundary}). \end{aligned}$$

There are many different ways to construct a group sequential boundary so that the

error probability constraints are satisfied according to

$$P_{\theta=0}(\text{Cross upper boundary before lower boundary}) = \alpha \quad (4.1)$$

$$P_{\theta=\delta}(\text{Cross lower boundary before upper boundary}) = \beta. \quad (4.2)$$

Pampallona and Tsiatis (1994) introduced a family of one-sided tests, indexed by a parameter Δ, where the degree of early stopping can be tailored to the investigator's needs by choosing an appropriate value for Δ. Pampallona et al. (2001) introduced further flexibility by defining spending functions for the type I and type II error probabilities. This contribution made it possible to deal with unpredictable group sizes and information levels without increasing the type I error probability. Jennison and Turnbull (2000, Chapter 7) found that the power remains close to its target value for such error spending designs, provided that deviations from the planned information sequence are moderate.

Let us now return to the example in the previous paragraph, and let us for a moment consider what can happen if a trial is not stopped according to the critical values. As for any $\theta \in \mathbf{R}$, there is under $\theta = 0$ a probability larger than zero to cross the lower boundary and then the upper boundary. We see in (4.1) that such sample paths are not included when calculating the overall type I error probability. When the type I error is calculated in this way, the presence of a futility boundary can thus pull down the upper boundary at later analyses. The critical value b_k required to stop is then smaller than what would have been the case without a futility boundary. Making use of the lost type I error probability associated with sample paths that cross the lower boundary before the upper boundary is sometimes referred to as re-claiming or buying back type I error.

The decision to stop a clinical trial is typically made after recommendation from a data monitoring committee (DMC). As discussed by Lan et al. (2003), it is however far from certain that a trial will be stopped, even though a boundary has been crossed. The experience of Proschan et al. (2006, Chapter 5) is that in practice, DMCs typically treat futility boundaries as more advisory than upper boundaries. There may be various reasons for a trial not being stopped, even though the futility boundary is crossed. There may for example be secondary variables that show promising results. Another reason could be that the decision about whether to stop for futility is influenced by other trials that are being run to investigate the same drug.

If type I error is re-claimed and the study may sometimes continue despite crossing the lower boundary, the overall type I error probability will be inflated. Chang and Chuang-Stein (2004) state that it should be possible to re-claim type I error. Snapinn (2006) takes a more cautious view and argues that while it is statistically valid to re-claim type I error, this should only be done if it is certain that the futility stopping rule

is always applied. Proschan et al. (2006, Chapter 5) go a step further and recommend making futility boundaries non-binding, so that the type I error is controlled even if the futility boundary is always overruled.

We are not aware of regulatory guidance that explicitly rules out re-claiming type I error, but there are documents that point in this direction. In regulatory guidance from the FDA (2006), it is assumed that the upper boundary is derived independently of the lower boundary, in a similar fashion to Proschan et al. (2006, Chapter 5). It can be argued that if there is no guarantee that the futility boundary will always be applied, it is preferable not to re-claim type I error. Another concern with binding futility boundaries is that it actually is possible to construct an arbitrarily small critical value b_K, by using a very aggressive futility boundary at previous analyses. Hence, it is possible to construct group sequential designs with binding futility boundaries that make it possible to reject the null hypothesis even though $Z_k < z_{1-\alpha}$. Burman and Sonesson (2006) have pointed to similar problems for adaptive designs.

In Section 4.2 we discuss existing group sequential designs with non-binding futility boundaries. As it is of interest to assess the efficiency of the designs in Section 4.2, we have developed a new method to derive optimal designs in the same class. The designs obtained from our optimisation procedure are a little conservative with respect to attained type I error, which is typical for group sequential designs with non-binding futility boundaries. Our optimal designs still maintain most of the benefits of early stopping found in group sequential designs with binding futility boundaries. The optimisation method is outlined in Section 4.3, which also contains an illustrative example. Further details of the optimisation method are provided in Section 4.6. Our approach builds on the method of dynamic programming, that was used in Chapter 2. It is however not simply an application of dynamic programming, since the non-binding futility boundary makes the optimisation problem more complex and difficult to solve.

The next step is to use our optimal designs to assess the efficiency of the designs in Section 4.2. In Section 4.4 we present efficiency comparisons between error spending designs, stochastic curtailment designs, and the optimal group sequential designs discussed in Section 4.3. Finally, we summarise our conclusions in Section 4.5.

4.2 Existing designs with non-binding futility boundaries

4.2.1 Formulation

The issues described in Section 4.1 raise the question of whether it is possible to enjoy the benefits of futility stopping, while making sure that the type I error is controlled. Setting $a_k = -\infty$, for $k = 1, \ldots, K-1$, makes sure that the type I error is controlled, but the benefits of futility stopping are then lost. A more interesting suggestion is

given by Proschan et al. (2006, Chapter 5), who advise against re-claiming type I error and propose making futility boundaries non-binding. They strongly prefer to give the DMC increased flexibility by deriving the upper boundary independently of the lower boundary. The type I error probability is then controlled even if the futility boundary is overruled, but slightly below its intended value if the study is in fact stopped for futility whenever the lower boundary is crossed. This proposal is thus a little conservative. It makes use of the fact that even though we would like the type I error probability to be exactly α, the most important thing is to make sure that it does not exceed α. Hence, we shall consider procedures with type I error probability at most α, with equality when the futility boundary is always overruled. We also require power of $1-\beta$ under $\theta = \delta$, assuming that the futility boundary is always applied. Instead of the constraints defined in (4.1) and (4.2), we now require

$$P_{\theta=0}(\text{Cross upper boundary ignoring lower boundary}) = \alpha \tag{4.3}$$

$$P_{\theta=\delta}(\text{Cross lower boundary before upper boundary}) = \beta. \tag{4.4}$$

We shall be referring to designs that satisfy (4.3) and (4.4) as non-binding designs, while designs that satisfy (4.1) and (4.2) will be referred to as binding designs.

The software package East-5 (2007) provides various ways to construct group sequential tests that satisfy (4.3) and (4.4). One possibility is to modify the power family tests of Pampallona and Tsiatis (1994) a little, making the boundary to reject H_0 slightly more conservative. We will however focus on error spending designs, as these have the additional benefit of being able to cope with unpredictable group sizes and information levels. Another possibility that we shall consider in Section 4.2.3 is to use methods for stochastic curtailment to derive a lower boundary.

4.2.2 Error spending designs

The construction of error spending designs that satisfy (4.3) and (4.4) bears similarities with the error spending designs with binding futility boundaries of Pampallona et al. (2001). We first calculate the information

$$\mathcal{I}_{fix} = \frac{(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta))^2}{\delta^2}$$

needed for power of $1-\beta$ at $\theta = \delta$ in a fixed sample trial. The next step is to define functions f and g that are used to spend the type I and type II error probabilities, respectively. These functions must be non-decreasing and satisfy $f(0) = g(0) = 0$, $f(\mathcal{I}) = \alpha$ and $g(\mathcal{I}) = \beta$, for $\mathcal{I} \geq \mathcal{I}_{max}$. One popular choice is the so-called ρ family, with f and g defined according to

$$f(\mathcal{I}) = \alpha \min\{(\mathcal{I}/\mathcal{I}_{max})^{\rho}, 1\} \tag{4.5}$$

$$g(\mathcal{I}) = \beta \min\{(\mathcal{I}/\mathcal{I}_{max})^{\rho}, 1\}, \tag{4.6}$$

where the parameter $\rho > 0$ and governs the amount of early stopping. For given ρ and $\mathcal{I}_{max}$, we have (4.5) and (4.6). The next step is to search for the correct $\mathcal{I}_{max}$ so that, for equally spaced analyses say, $a_K = b_K$ at the final analysis. When $\mathcal{I}_{max}$ has been calculated, the inflation factor can be calculated as $R = \mathcal{I}_{max}/\mathcal{I}_{fix}$.

Suppose that interim analyses are performed at information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$. We can then define $\pi_{1,1} = f(\mathcal{I}_1)$, $\pi_{2,1} = g(\mathcal{I}_1)$ and, for $k = 2, \ldots, K$,

$$\pi_{1,k} = f(\mathcal{I}_k) - f(\mathcal{I}_{k-1})$$

and

$$\pi_{2,k} = g(\mathcal{I}_k) - g(\mathcal{I}_{k-1}).$$

The critical values a_k and b_k at analysis k can be calculated in turn, starting with $k = 1$ and working upwards, as the solutions to

$$\pi_{1,k} = P_{\theta=0}(Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1},\, Z_k \geq b_k)$$

and

$$\pi_{2,k} = P_{\theta=\delta}(a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1},\, Z_k \leq a_k).$$

The upper boundary is thus calculated without taking the lower boundary into account. The upper boundary is however taken into account when constructing the lower boundary, so that the power is at least $1 - \beta$ at $\theta = \delta$, with equality if the study is always stopped when crossing the lower boundary. For given K, α, β, ρ and planned information sequence $\mathcal{I}_1, \ldots, \mathcal{I}_{max}$, we can thus calculate the critical values $a_1, b_1, \ldots, a_K, b_K$ that define a one-sided group sequential test satisfying the error probability constraints defined in (4.3) and (4.4). Several iterations, where $\mathcal{I}_{max}$ is adjusted, may be necessary to obtain $a_K = b_K$ at the final analysis K.

4.2.3 Designs based on stochastic curtailment

Suppose that in a K-group sequential design to test $H_0 : \theta \leq 0$ versus the alternative $H_1 : \theta > 0$, an upper boundary $b_1, \ldots, b_K$ that satisfies (4.3) has been chosen. This upper boundary may have been derived from an alpha spending function or some other approach. Without a lower boundary, this upper boundary in itself defines a one-sided group sequential test, with no possibility of early stopping for futility. Given the upper boundary, we can, assuming equally spaced analyses or some other pattern

for the interim analyses, derive the maximum information level $\mathcal{I}_{max} = R \times \mathcal{I}_{fix}$ that gives power $1 - \beta$ at $\theta = \delta$. In the context of methods for stochastic curtailment, Jennison and Turnbull (2000, Chapter 10) refer to this test without a lower boundary as a reference test.

Suppose that we are happy with the upper boundary of the reference test, but would also like to enjoy the benefits of early stopping for futility. Instead of deriving the lower boundary through a beta spending function, we shall focus on another approach, called stochastic curtailment. In stochastic curtailment methods, the trial is stopped for futility when the probability of rejecting $H_0 : \theta \leq 0$ later on in the trial falls below a certain threshold.

It is important to note that there are different ways to define the probability that the reference test rejects H_0 later on in the trial. The reason is that it is not obvious which value of the treatment effect θ to use in the calculation. In this section we will give a precise definition for three possibilities, which in turn give three different criteria for futility stopping:

- predictive power calculated given the current posterior distribution for θ, where the posterior distribution given $Z_k = z_k$ is calculated for the prior distribution $\pi(\theta)$,
- conditional power calculated for $\theta = \hat{\theta}$, where $\hat{\theta}$ is the maximum likelihood estimate of θ,
- conditional power calculated for $\theta = \delta$.

Various authors have described methods for stochastic curtailment, including the three methods that we shall focus on. Lachin (2005) gives an overview of methods for futility stopping based on conditional power. Jennison and Turnbull (2000, Chapter 10) describe different methods for stochastic curtailment, including conditional power, predictive power and a parameter-free approach. From a Bayesian point of view, Spiegelhalter et al. (1986) argue that the use of predictive power may be more appropriate than conditional power. Proschan et al. (2006, Chapter 3) discuss further possibilities, for example using updated estimates of nuisance parameters in the conditional power calculation.

Let us now return to the reference test with an upper boundary that satisfies (4.3). Given the upper boundary of the reference test we can, after having observed $Z_k = z_k$ at interim analysis k, calculate

$$\begin{aligned}
\mathrm{PP}(z_k) &= \int d\theta \pi^{(k)}(\theta|z_k)\mathrm{CP}_\theta(z_k) && (4.7)\\
\mathrm{CP}_\delta(z_k) &= P_\delta(\text{Cross upper boundary at analyses } k+1,\ldots,K|Z_k = z_k). && (4.8)\\
\mathrm{CP}_{\hat{\theta}}(z_k) &= P_{\hat{\theta}}(\text{Cross upper boundary at analyses } k+1,\ldots,K|Z_k = z_k) && (4.9)
\end{aligned}$$

For (4.7), $\pi^{(k)}(\theta|z_k)$ will be calculated for the improper prior $\pi(\theta) = 1$. This prior takes the value 1 over the whole real line and is called improper because the integral of the pdf is not finite. If the treatment effect is estimated to be $\hat{\theta}_k$ at analysis k, with information level $\mathcal{I}_k$, the posterior distribution for θ is $N(\hat{\theta}_k, \mathcal{I}_k^{-1})$. Another possibility would be to assume that the prior distribution for θ follows a normal distribution. We refer to Proschan et al. (2006, Chapter 3) for a discussion of how to choose the mean and variance of the prior distribution in this case.

It is noteworthy that since (4.8) is calculated under the alternative $\theta = \delta$, it can be viewed as a special case of (4.7), with a one-point prior at $\theta = \delta$. It is on the other hand not possible to choose a prior so that (4.9) becomes a special case of (4.7). In Section 4.4.2 we will see that futility stopping based on (4.9) is not very efficient. The fact that there is no prior distribution that leads to a one-point posterior distribution at $\theta = \hat{\theta}$, may be a reason for the inefficiency.

If either (4.8), (4.9) or (4.7) is below a certain threshold, this may indicate that a trial is unlikely to achieve a statistically significant result. We shall be focusing on stopping rules of the type

$$\text{Stop for futility at analysis } k \text{ if } \mathrm{PP}(z_k) \leq \eta_1 \tag{4.10}$$

$$\text{Stop for futility at analysis } k \text{ if } \mathrm{CP}_\delta(z_k) \leq \eta_2 \tag{4.11}$$

$$\text{Stop for futility at analysis } k \text{ if } \mathrm{CP}_{\hat{\theta}}(z_k) \leq \eta_3. \tag{4.12}$$

If a criterion of this type is used at each interim analysis of a clinical trial, it is straightforward to derive an equivalent rule, where $Z_k = z_k$ is compared to a lower boundary $a_1, \ldots, a_{K-1}$. Combined with an upper boundary $b_1, \ldots, b_K$ that satisfies (4.3) and information levels $\mathcal{I}_1, \ldots, \mathcal{I}_{max}$, we can calculate the power at a certain effect size δ. The presence of the futility boundary will imply some loss of power compared to the reference test. If it is important to satisfy the power requirement, a reference test with power higher than $1 - \beta$ must be chosen.

4.3 Optimal group sequential designs with non-binding futility boundaries

4.3.1 Motivation

Barber and Jennison (2002) found that for group sequential designs with binding futility boundaries, ρ family error spending designs are close to optimal in a wide range of situations. Moreover, error spending designs have other advantages, for example the ability to cope with unpredictable information sequences without inflating the type I error probability. It is thus of interest to assess whether ρ family error spending

designs with non-binding futility boundaries are also close to optimal in their class. If the efficiency loss compared to optimal designs is small, error spending desígns would provide an efficient class of designs that are convenient for use in practice.

The methods for stochastic curtailment described in Section 4.2.3 also have appealing features. The focus of these methods is primarily on the lower boundary and stopping for futility, but it is also of interest to assess the overall efficiency of these designs compared to optimal designs. It may well be that we can find stochastic curtailment designs that are close to optimal, providing an efficient alternative to error spending designs. An additional benefit of these methods is that they are easy to communicate. We believe that the probability of rejecting H_0, given current data, might be easier to interpret than a critical value on the Z scale.

4.3.2 Formulating the optimality criteria

While the designs in Section 4.2 satisfy the error probability constraints (4.3) and (4.4), we have not yet assessed their efficiency. Efficiency will be expressed in terms of expected information, which can easily be converted to sample size. For normally distributed data with known variance we have $\mathcal{I}_k = n_k/(2\sigma^2)$, where n_k is the cumulative per-treatment sample size at analysis k. For a K-group sequential design with critical values $a_1, b_1, \ldots, a_K, b_K$, we define the expected information, at a given value of the effect size θ, as

$$E_\theta(\mathcal{I}) \;=\; \sum_{k=1}^{K} P_\theta\{Z_1 \in C_1, \ldots, Z_{k-1} \in C_{k-1}, Z_k \le a_k \text{ or } Z_k \ge b_k\}\mathcal{I}_k, \quad (4.13)$$

where $C_k = (a_k, b_k)$. The expected information is thus calculated assuming that the futility boundary will be strictly adhered to, reflecting how the investigator expects the trial to be monitored. This would appear to be the most logical definition, as the constraint on how to calculate the type I error probability can be viewed as an additional requirement imposed by regulatory authorities. Rather than focusing on one value of θ, a more appropriate way to assess the performance of the designs in Section 4.2 may be to calculate (4.13) across a range of values of θ. We shall present a method to derive group sequential designs that are optimal, in the sense that they minimise

$$\tilde{F} = \frac{E_0(\mathcal{I}) + E_{\delta/2}(\mathcal{I}) + E_\delta(\mathcal{I})}{3} \quad (4.14)$$

subject to the constraints in (4.3) and (4.4). Thereafter, we will assess the performance of designs in Section 4.2 that satisfy the same error probability constraints, by calculating $\tilde{F}$ and comparing this number to that obtained for the optimal design. The expected information in (4.14) is calculated according to (4.13), i.e. assuming that

the futility boundary will always be applied. We note that if desirable, it would be straightforward to use our method to optimise another weighted average than (4.14).

4.3.3 Derivation of optimal designs

The first step necessary to enable an efficiency comparison with the designs in Section 4.2 is to develop a method to derive optimal designs. For binding futility boundaries, Barber and Jennison (2002) describe how dynamic programming can be used to derive one-sided group sequential tests that are optimal, in the sense that they minimise the expected sample size while satisfying (4.1) and (4.2). This is done by defining a prior for θ, a cost of sampling and a loss function that defines costs for making the wrong decision about θ. Each interim analysis can be thought of as a decision node. At each decision node, the decisions are based on expected future costs under the current posterior distribution for θ.

Suppose that instead we wish to minimise $\tilde{F}$ as defined in (4.14), subject to the error probability constraints in (4.3) and (4.4). We can follow an approach that bears similarities with the method of Barber and Jennison (2002), but which also has important differences. Consider the unconstrained optimisation problem of finding the critical values that minimise

$$\tilde{F} + \lambda_1 P_1 + \lambda_2 P_2, \tag{4.15}$$

where λ_1 and λ_2 are positive and P_1 and P_2 are the probabilities, defined in (4.3) and (4.4), of rejecting and accepting H_0. Provided that we can find the solution $a_1, b_1, \ldots, a_K, b_K$ that minimises (4.15), it remains to choose the Lagrange multipliers λ_1 and λ_2 so that the solution gives $P_1 = \alpha$ and $P_2 = \beta$. The design that minimises (4.15) for these values of λ_1 and λ_2 must then also minimise $\tilde{F}$ among all designs satisfying the same constraints.

P_1 and P_2 are however probabilities under two different stopping rules, so this is not a standard Bayes sequential decision problem. Hence, it is not straightforward to solve (4.15) through the method dynamic programming used in Section 2.2.2. The key ingredient that makes this problem different is that the optimal critical values at analysis k depend on the critical values at both previous and future analyses. A slightly different approach is thus warranted, details of which are provided in Section 4.6.

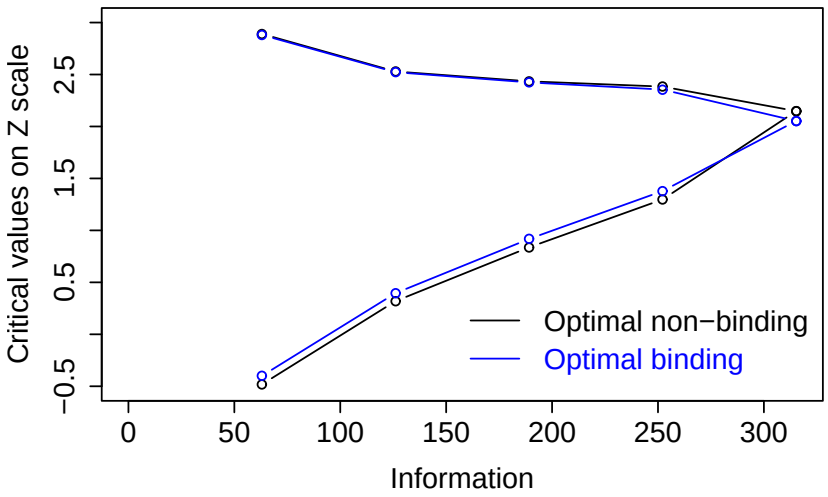

Figure 4-1: Critical values for optimal group sequential designs with non-binding (black solid line) and binding (blue solid line) futility boundaries

4.3.4 An illustrative example

Consider a 5-group sequential design with type I error probability $\alpha = 0.025$ and 90% power at $\delta = 0.2$. The information required for a fixed sample test is

$$\mathcal{I}_{fix} = \frac{(\Phi^{-1}(1-0.025) + \Phi^{-1}(1-0.1))^2}{0.2^2} = 263.$$

For normally distributed data with known variance, this can be converted to patients per treatment group by using the identity $n_{fix} = 2\sigma^2\mathcal{I}_{fix}$. Assuming equally spaced analyses and an inflation factor $R = 1.2$, we schedule interim analyses at

$$\mathcal{I}_k = \mathcal{I}_{max}\frac{k}{K},\ k = 1, \ldots, K,$$

where $\mathcal{I}_{max} = R \times \mathcal{I}_{fix} = 315$. Optimising the design for $\tilde{F}$, subject to the constraints in (4.3) and (4.4), gives the boundaries displayed with a black solid line in Figure 4-1. For comparison we also include the boundaries (blue solid line) for the design with binding futility boundaries, which has also been optimised for $\tilde{F}$. The boundaries of the two designs are very similar, but we can observe that the critical values are slightly more conservative for the design with non-binding futility boundaries. The binding design can stop for efficacy for lower Z values since, as discussed in Section 4.1, treating the futility boundary as binding makes it possible to re-claim type I error.

Figure 4-2 shows four power curves that merit some comment. Two of the curves are

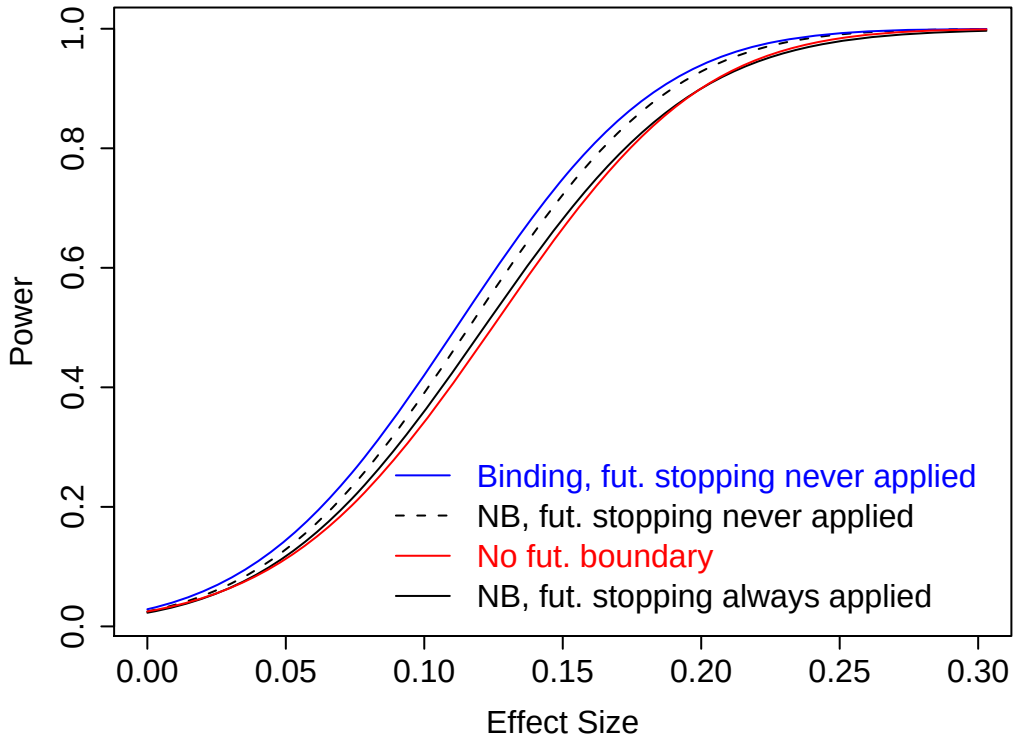

Figure 4-2: Power for optimal design with non-binding futility boundaries, if futility boundary is always applied (black solid line) and never applied (black dashed line). Also shown are the power curve (red line) for a design with no futility boundary and optimal upper boundary and the power curve (blue solid line) for an optimal binding design for which the futility boundary is never applied.

the result of two different ways of applying the optimal non-binding boundary displayed with solid line in Figure 4-1. The solid line displays the probability of rejecting H_0 if the futility boundary is strictly adhered to. Hence, the probability of rejecting H_0 at $\theta = 0$ is 0.023, while the probability of rejecting H_0 at $\theta = 0.2$ equals 0.90. From the point of view of the trial sponsor, this shows that the power requirement is satisfied if, as intended, the futility bound is always applied. The dashed line shows what happens to the power curve of the design with a non-binding futility boundary, if the futility boundary is never applied. The type I error is controlled at $\alpha = 0.025$ also in this case, which should be reassuring to regulators. The power is also slightly higher, 0.93 rather than 0.90, since sample paths that first cross the lower boundary and later the upper boundary are now allowed to continue.

We have seen that non-binding futility boundaries make designs a little conservative with respect to type I error. Using designs without a futility boundary is an obvious

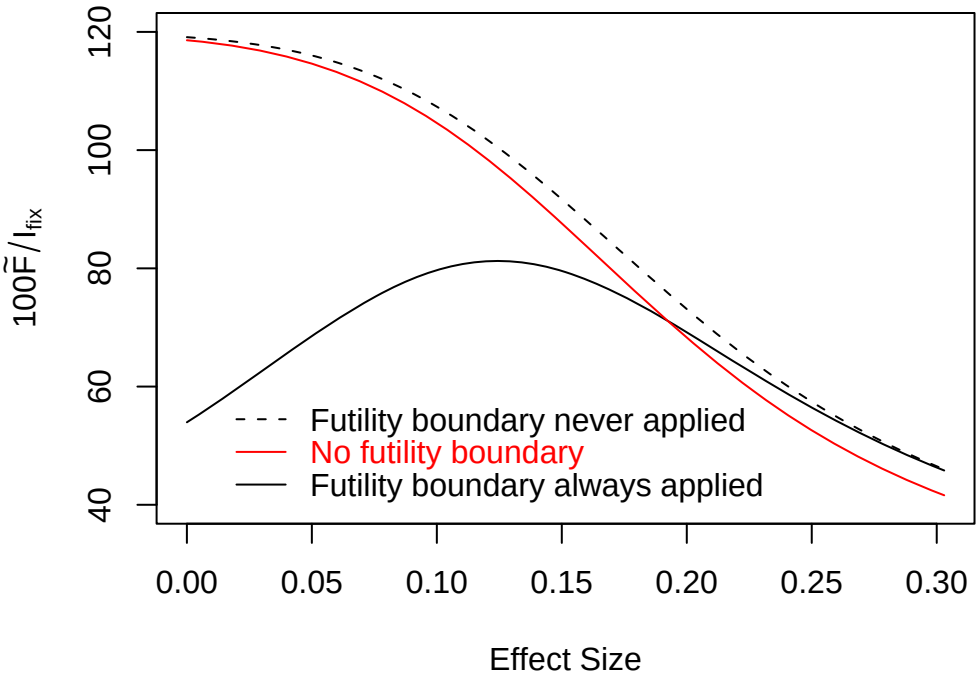

Figure 4-3: $100\tilde{F}/\mathcal{I}_{fix}$ for optimal non-binding design if futility boundary is always applied (black solid line) and never applied (black dashed line). Also shown is $100\tilde{F}/\mathcal{I}_{fix}$ for a design with no futility boundary and optimal upper boundary (red solid line).

way to get around the problem of how to treat the lower boundary. Such designs have type I error α and power $1-\beta$ at $\theta = \delta$, but do not enjoy the benefits of futility stopping. The power curve for such a design, with optimised upper boundary $b_1, \ldots, b_5$, but no futility boundary, is shown with a solid red line in Figure 4-2. The power curve is very close to the solid line, i.e. the design with a futility boundary that is always applied. Finally, we also show a curve (solid blue line) for the optimal binding design in Figure 4-1, for the case when the futility boundary is never applied. This design has type I error 0.029 and power 0.94 at $\theta = 0.2$, giving an indication of what can happen if a binding futility boundary is incorrectly ignored.

Figure 4-3 shows the ratio $100\tilde{F}/\mathcal{I}_{fix}$ for the three different ways of applying the futility boundary. The black solid line shows $100\tilde{F}/\mathcal{I}_{fix}$ calculated according to (4.13), i.e. by assuming that the futility boundary will be strictly adhered to. As we would expect, the presence of a futility boundary is crucial for achieving good efficiency for small values of θ. It is however worth remembering that applying the futility boundary

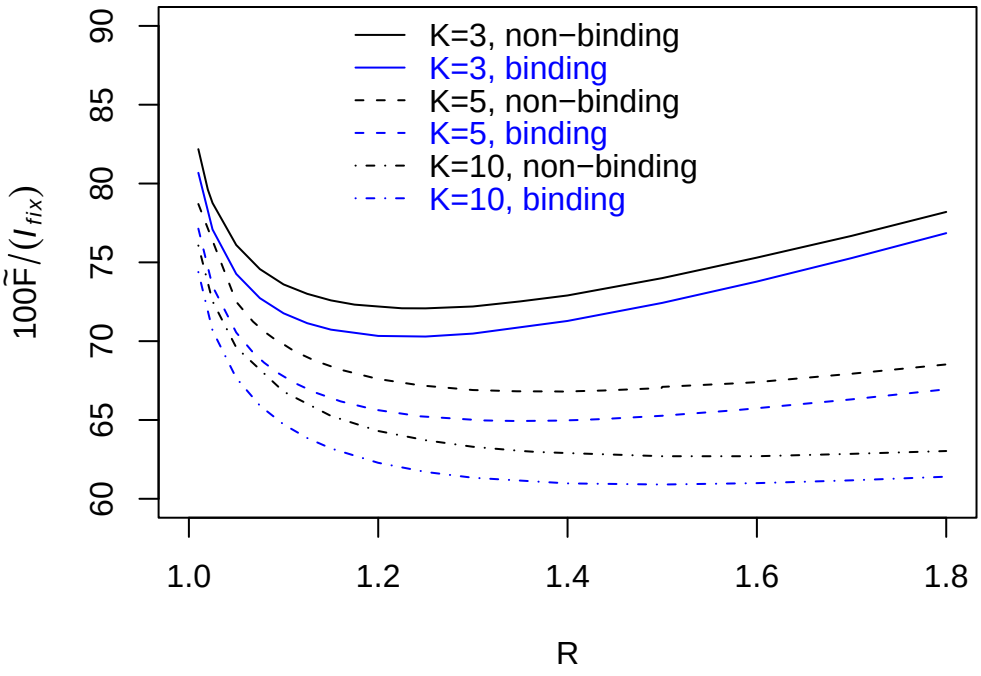

Figure 4-4: $100\tilde{F}/\mathcal{I}_{fix}$ for optimal K-group sequential designs with equally spaced analyses and non-binding futility boundaries. The designs have inflation factor R, type I error $\alpha = 0.025$ and power $1-\beta$ at $\theta = \delta$.

also implies a loss of power, as illustrated in Figure 4-2. Also shown in Figure 4-3, with red solid line, is $100\tilde{F}/\mathcal{I}_{fix}$ for a design with optimised upper boundary $b_1, \ldots, b_5$, but without a futility boundary. The design without futility boundary shows good efficiency for high values of θ. It is on the other hand very inefficient for small values of θ. The results are similar for the design, displayed with dashed line, with a futility boundary that is not applied, with slightly higher $\tilde{F}$ and slightly higher power. We conclude that if values of θ close to the null hypothesis are of interest, designs with futility boundaries are to be preferred. These can be non-binding if that is what is required. Non-binding designs are a little conservative with respect to type I error, but show good efficiency across a range of values of θ. If the futility boundary is always applied, this class of designs can deliver substantial efficiency gains, compared to the corresponding fixed sample trial.

Figure 4-4 shows the efficiency gains of optimal K-group sequential designs with non-binding futility boundaries and inflation factor R, compared to the corresponding fixed sample designs. As will be the case for the remainder of this chapter, $100\tilde{F}/\mathcal{I}_{fix}$

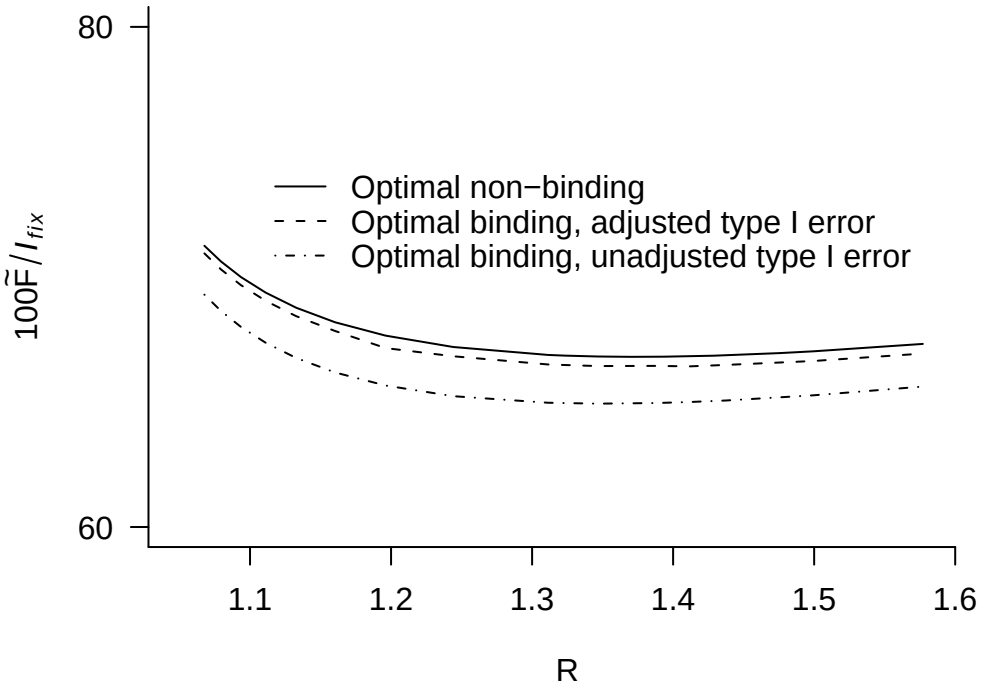

Figure 4-5: Efficiency comparison between three 5-group sequential designs, with equally spaced analyses, that are optimal with respect to $\tilde{F}$ within their respective class.

is calculated assuming that the futility boundary is always applied. For each K, the ratio $100\tilde{F}/\mathcal{I}_{fix}$ initially decreases in R, but as R increases there comes a point where $\tilde{F}/\mathcal{I}_{fix}$ starts to increase.

We also show the efficiency of K-group binding designs with varying inflation factor, that have been optimised with respect to $\tilde{F}$. A similar pattern, with an initial decrease of $100\tilde{F}/\mathcal{I}_{fix}$ with increasing R, followed by an increase, is seen for the binding designs. For given K, the optimal efficiency is however achieved for a slightly smaller R than for non-binding designs. For fixed R and K, the efficiency loss of the non-binding designs compared to the binding is about 2%, which can be interpreted as the cost incurred for the additonal constraint that is imposed on the type I error.

Figure 4-5 shows, with solid line, $100\tilde{F}/\mathcal{I}_{fix}$ for optimal 5-group designs with non-binding futility boundaries. As a benchmark, two other designs have also been included in the figures. We consider optimal group sequential designs with binding futility boundaries, with adjusted and unadjusted α levels. In summary, the following three classes of designs are displayed in Figures 4-5:

1. Optimal group sequential designs with non-binding futility boundaries, displayed with a solid line.

2. Optimal group sequential designs with binding futility boundaries and adjusted type I error, displayed with a dashed line. Here, the type I error has been adjusted downwards, so that for the fixed K and R, it equals the type I error attained for the optimal design with non-binding futility boundary.

3. Optimal group sequential designs with binding futility boundaries and unadjusted type I error, displayed with dot-dashed line. These designs have type I error probability α, calculated according to (4.1).

The designs with unadjusted type I error ought to give lower expected sample size than the other designs, as their attained type I error probabilities are higher. This is confirmed in Figure 4-4 and Figure 4-5. The designs with binding futility boundaries and adjusted type I error perform slightly better than the design with non-binding futility boundary satisfying the same error probability constraints. This should be the case, as we are comparing the optimal design within a certain class to a design that happens to be in this class, but is not necessarily optimal. The difference is however very small. Hence, most of the efficiency loss associated with the design with non-binding futility boundary is due to the differences in attained type I error, as a result of the futility boundary being treated differently.

4.4 Efficiency comparisons with existing designs

4.4.1 Assessment of error spending designs

We now consider error spending designs with type I error according to (4.3) and type II error according to (4.4). Hence, the type I error is calculated without taking the futility boundary into account, while the power is calculated assuming that the futility boundary is always applied. To enable a fair comparison between ρ family error spending designs and optimal group sequential designs, we consider optimal group sequential designs with the same type I and type II errors and cumulative group sizes as the corresponding error spending designs. The efficiency of the designs is evaluated by calculating $\tilde{F}$, i.e. by assuming that the study is stopped for futility whenever the lower boundary is crossed.

For K equally spaced analyses there is, for α and β held fixed, a one-to-one correspondence between the choice of ρ and the inflation factor R necessary for power $1-\beta$ at $\theta = \delta$. When $K = 5$, we obtain $R = 1.07$ for $\rho = 3$ and $R = 1.31$ for $\rho = 1$. For each value of ρ we obtain a maximum information level $\mathcal{I}_{max}$, and given our assumption about equally spaced analyses, an information sequence $\mathcal{I}_1, \ldots, \mathcal{I}_{max}$.

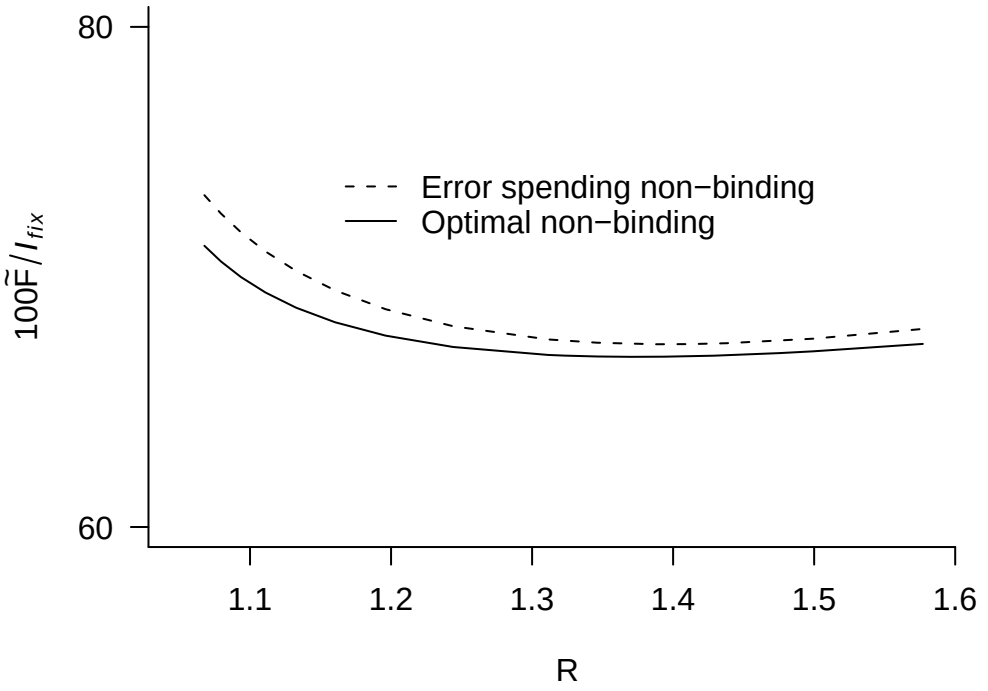

Figure 4-6: Efficiency comparison between $K = 5$, ρ family error spending designs and corresponding optimal 5-group sequential designs. All designs have non-binding futility boundaries, equally spaced analyses, type I error α and power $1 - \beta$ at $\theta = \delta$.

Given this information sequence, we can find the design that is optimal with respect to $\tilde{F}$ and compare its efficiency to that of the error spending design. Figure 4-6 shows an efficiency comparison between ρ family, $K = 5$ group, error spending designs and optimal designs, where we have drawn R instead of ρ on the horizontal axis. The efficiency loss of the error spending designs is small, typically between $0.5\% - 2\%$. Hence, we can be confident about using non-binding error spending designs in clinical trials, without losing much efficiency compared to optimal non-binding designs. This is an important conclusion, and was in Section 4.3.1 part of our motivation for developing a method to derive optimal designs.

Figure 4-7 displays comparisons of non-binding, $\rho = 1$ error spending designs, with the corresponding optimal non-binding designs, for different values of K. Also shown are the optimal binding designs with adjusted and unadjusted type I error, of the type displayed in Figure 4-5. We see a similar pattern for $K = 3$ and $K = 10$ as shown for $K = 5$, in Figures 4-5 and 4-6.

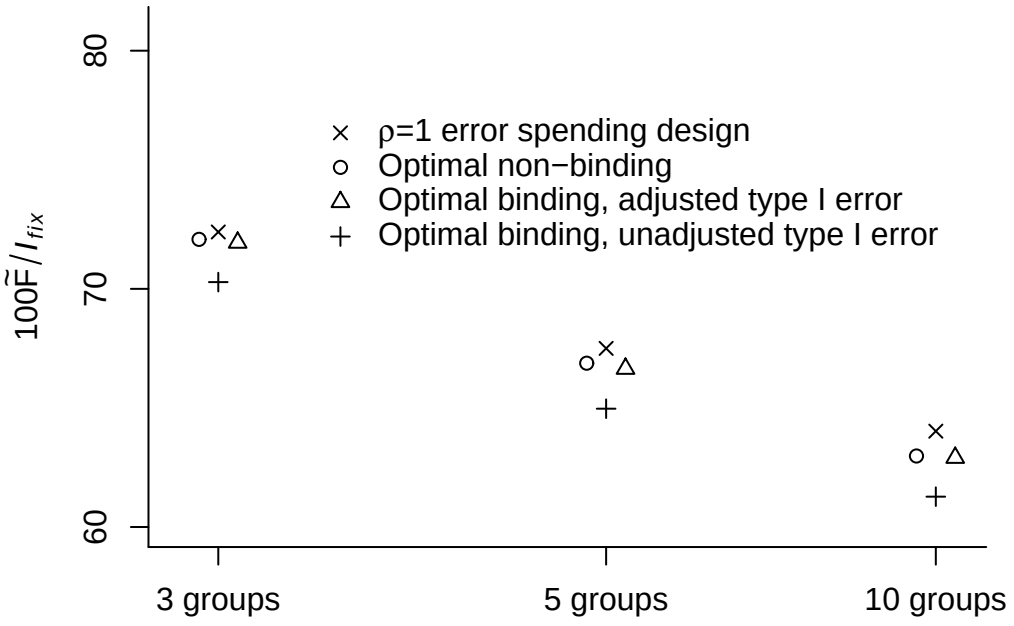

Figure 4-7: Efficiency comparison between K-group, $\rho = 1$ error spending designs and corresponding optimal K-group sequential designs. All designs have non-binding futility boundaries, equally spaced analyses, type I error α and power $1 - \beta$ at $\theta = \delta$.

4.4.2 Assessment of methods for stochastic curtailment

We shall now assess the efficiency of methods based on stochastic curtailment, by comparing the efficiency of such designs to that of optimal designs derived using our optimisation method described in Section 4.3. We will first consider an example of how a stochastic curtailment design can be constructed in practice. Thereafter, efficiency comparisons, of the type presented for error spending designs in Section 4.4.1, will be made. We will describe a design with futility stopping based on predictive power, but our description of how to derive the lower boundary is applicable also for the other methods described in Section 4.2.3.

Consider a group sequential test of $H_0 : \theta \leq 0$ against the alternative $H_1 : \theta > 0$, with type I error α and power $1 - \beta$ at $\theta = \delta$. Suppose that it has been decided to have $K = 5$ equally spaced analyses with an upper boundary $b_1, \ldots, b_K$ that satisfies (4.1). This upper boundary can play the role of a reference test, and we can search for the maximum information $\mathcal{I}_{max} = R \times \mathcal{I}_{fix}$ that is required to satisfy the power

requirements, without a futility boundary present. The reference test thus has interim analyses at

$$\mathcal{I}_k = \mathcal{I}_{max} \times \frac{k}{K}, \quad k = 1, \ldots, K,$$

with early stopping to reject H_0 if $\hat{\theta}_k \sqrt{\mathcal{I}_k} \geq b_k$.

Let us now consider how to add a lower boundary to this design by using stochastic curtailment based on predictive power. The lower boundary is constructed by using a rule of the type defined in (4.10). At each interim analysis, we can calculate the predictive power to reject the the null hypothesis of the reference test. Suppose that the trial is stopped for futility if the predictive power falls below η_1. At each interim analysis, we can search for the critical value a_k where the predictive power to reject the null hypothesis of the reference test equals η_1. For the new group sequential design, with lower boundary $a_1, \ldots, a_K$ and the upper boundary of the reference test, we can calculate properties such as power and expected sample size.

We shall now consider the efficiency of stochastic curtailment designs compared to optimal group sequential designs. To this end, the stochastic curtailment designs will be given the same upper boundary as the optimal designs, while we will assess the efficiency of the lower boundaries derived from stochastic curtailment methods. To enable a fair comparison between two designs, it is desirable that they satisfy the same error probability constraints and have interim analyses at the same cumulative sample sizes. It follows from (4.3) that two designs with the same upper boundary have the same type I error, as this is assessed without taking the futility boundary into account. The power curve will however be affected by the lower boundary. It is clear that for the stochastic curtailment design, the power is monotonically decreasing in η_1, the threshold that is used to decide whether to stop for too low predictive power. To make sure that both designs have power $1 - \beta$ at $\theta = \delta$, we can for the stochastic curtailment design perform a one-dimensional search for the threshold η_1 that gives the same power as the optimal design. For $K = 2$, any stochastic curtailment design derived in this way is equivalent to the optimal design, as the boundary point a_1 must be set to satisfy the requirement of power $1 - \beta$ at $\theta = \delta$. For $K > 2$, an additional degree of freedom is added for each interim analysis.

Consider two of the optimal designs that were used in Figure 4-6, for comparisons against $\rho = 1$ and $\rho = 3$ error spending designs. These designs have inflation factors of about 1.31 and 1.07, respectively. The design with an inflation factor of 1.31 might be considered aggressive and has an upper boundary that is broadly similar to the designs of Pocock (1977), with a substantial probability of early stopping. The design with inflation factor of 1.07 is more conservative with wider boundaries early on, similar to the O'Brien and Fleming (1979) designs. We consider group sequential designs with optimised lower and upper boundary. We now wish to investigate the merits of

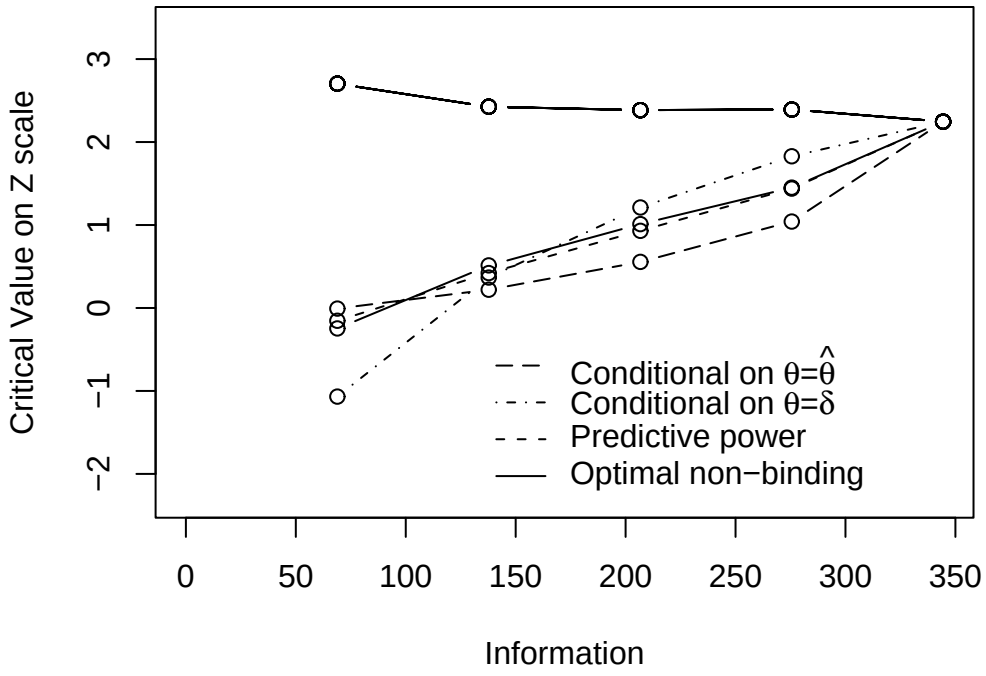

Figure 4-8: Critical values for optimal non-binding design and for non-binding designs with optimised upper boundary and lower boundary based on stochastic curtailment. All designs have $K = 5$ equally spaced analyses, inflation factor $R = 1.31$, type I error $\alpha = 0.025$ and power $1 - \beta$ at $\theta = 0.2$.

stochastic curtailment methods through an efficiency comparison with these optimised designs. We assess if it can still be highly efficient to use the same upper boundary as the optimal design, but construct a new lower boundary through stochastic curtailment.

Figures 4-8 and 4-9 show the critical values for the two optimal designs with $R = 1.07$ and $R = 1.31$, as well as for the different methods for stochastic curtailment defined in Section 4.2.3. We note that the rule based on predictive power is rather close to the optimal boundary. Compared to the optimal design, the rule based on conditional power at $\theta = \delta$ is a little more conservative early on, while the rule based on conditional power at $\theta = \hat{\theta}$ gives a more aggressive boundary at earlier interim analyses. In Figures 4-8 and 4-9 we use constant thresholds η_1, η_2 and η_3, to define the lower boundaries that are based on stochastic curtailment. Another possibility, not considered here, would be to vary the threshold in some systematic way over the interim analyses. The values of the thresholds are driven by our requirement that the designs should have 90% power at $\theta = \delta$. For $R = 1.31$, we have $\eta_1 = 0.10$, $\eta_2 = 0.61$ and

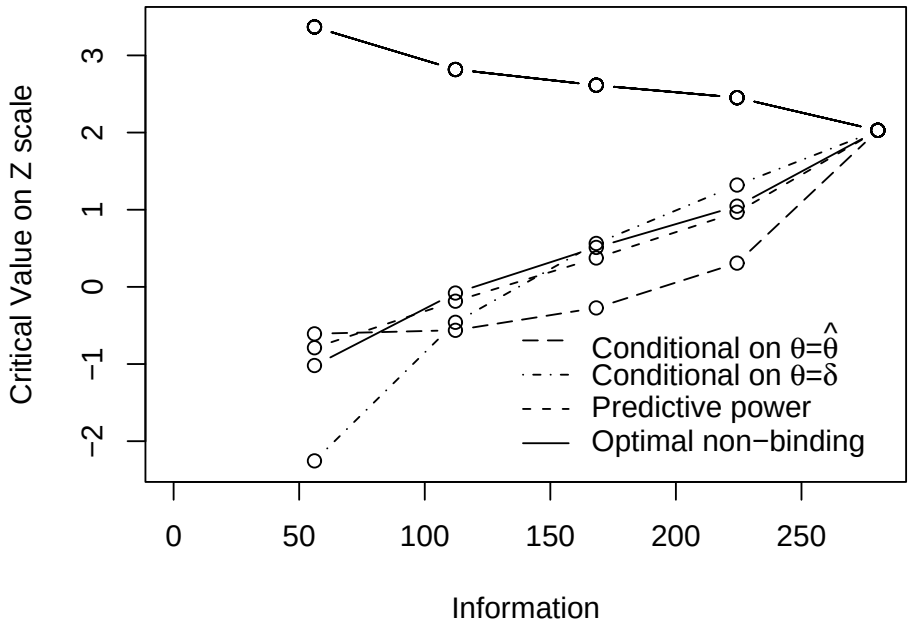

Figure 4-9: Critical values for optimal non-binding design and for non-binding designs with optimised upper boundary and lower boundary based on stochastic curtailment. All designs have $K = 5$ equally spaced analyses, inflation factor $R = 1.07$, type I error $\alpha = 0.025$ and power $1 - \beta$ at $\theta = 0.2$.

$\eta_3 = 0.0078$, while for $R = 1.07$, $\eta_1 = 0.029$, $\eta_2 = 0.35$ and $\eta_3 = 0.00008$. A different threshold can be obtained by finding the optimal design for a different inflation factor R, where a higher inflation factor will give a higher threshold. It may well be the case that it is found desirable to have a higher or lower threshold than the ones we report here. Proschan et al. (2006, Chapter 3) comment that a typical rule may be to stop if $\mathrm{CP}_\delta(z_k)$ is below $0.10 - 0.15$, i.e a little more conservative than the rules that we have considered. In our framework, we can obtain a rule with a value of η_2 in this range by setting $R \approx 1.01$. Because of their higher efficiency, we believe that the designs that we have considered, with a higher inflation factor, are to be preferred.

Figures 4-10 and 4-11 show the expected sample size functions for the optimal designs and for the rules based on stochastic curtailment for $R = 1.31$ and $R = 1.07$, respectively. The design with futility stopping based on predictive power is close to optimal across a range of values of θ. Conditional power based on $\theta = \delta$ also does rather well, but shows some inefficiency as we move away from $\theta = \delta$. We note that the method assumes $\theta = \delta$ also when the data observed so far do not support this

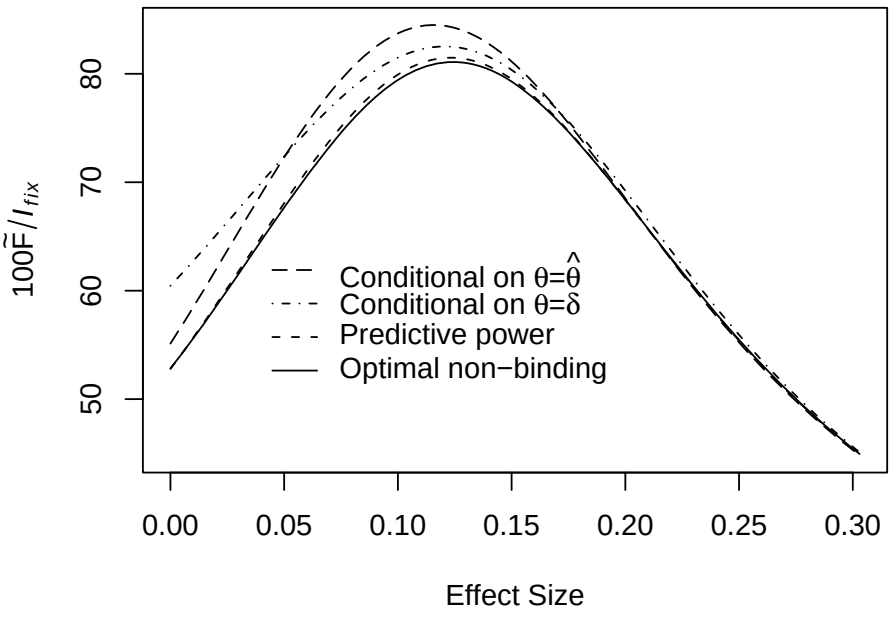

Figure 4-10: $100\tilde{F}/\mathcal{I}_{fix}$ for optimal non-binding design and for non-binding designs with optimised upper boundary and lower boundary based on stochastic curtailment. All designs have $K = 5$ equally spaced analyses, inflation factor $R = 1.31$, type I error $\alpha = 0.025$ and power $1 - \beta$ at $\theta = 0.2$.

assumption. To stop for futility at the first interim analysis, data have to be more extreme in the negative direction, than for the other methods. Conditional power based on $\hat{\theta}$ on the other hand seems to be a rather inefficient stopping rule. This rule relies on estimates of $\hat{\theta}$, without taking into account the variability of the estimates. This might be a source of inefficiency, in particular early on in the trial when estimates are uncertain. The boundaries based on predictive power are somewhere in between those based on $\theta = \hat{\theta}$ and $\theta = \delta$, and are the closest to the optimal boundary. Hence, we would recommend futility stopping based on predictive power for a constant threshold η_1 as an efficient stopping rule, that should be easy to implement in practice. The value of η_1 can be chosen by considering the power curve and the probability of early stopping that is achieved for a given threshold.

4.5 Discussion

We have described issues with how to calculate the type I error in one-sided group sequential designs with futility stopping. These issues are related to the conduct of

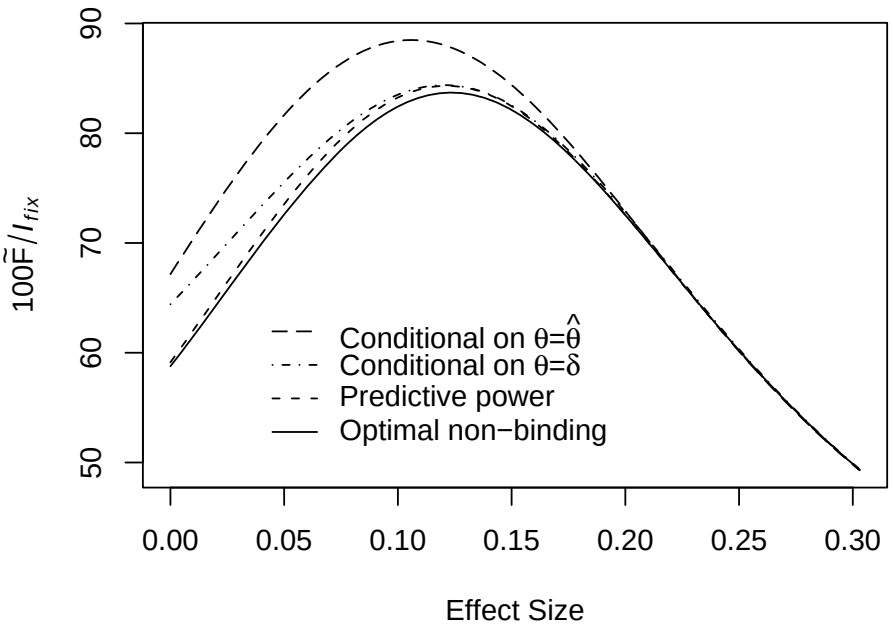

Figure 4-11: $100\tilde{F}/\mathcal{I}_{fix}$ for optimal non-binding design and for non-binding designs with optimised upper boundary and lower boundary based on stochastic curtailment. All designs have $K = 5$ equally spaced analyses, inflation factor $R = 1.07$, type I error $\alpha = 0.025$ and power $1 - \beta$ at $\theta = 0.2$.

group sequential designs and the fact that it is challenging to prove that a futility boundary will always be applied. Several authors have pointed out that in clinical trial practice, it may be difficult to justify the use of binding futility boundaries and some of them strongly advice against re-claiming type I error. The simplest solution is to remove the futility boundary from the design, but retain the possibility of early stopping for benefit. This possibility was considered in Section 4.3.4, and can still give good efficacy for high values of θ. The drawback is that such a method is very inefficient if the treatment effect is close to the null hypothesis. It can also be argued that there are ethical reasons to stop, if it is unlikely that the trial will give a positive result.

A better solution is to require that the probability of crossing the upper boundary equals the desired type I error probability when the lower boundary is not taken into account. This points to group sequential designs with non-binding futility boundaries of the type recommended by Proschan et al. (2006, Chapter 5). Using non-binding designs results in a small loss of power, but makes sure that the type I error probability is at most α. This should be reassuring to regulatory authorities, who might question that a binding futility boundary will always be applied. This chapter has dealt with the

efficiency of such designs, with a focus on error spending designs and designs based on stochastic curtailment. We have found error spending designs to give efficient designs. Good efficiency was also obtained for designs with optimised upper boundary and a lower boundary defined by the predictive power being equal to a certain threshold. It turned out to be surprisingly efficient to let this threshold be constant across the interim analyses. The other methods for stochastic curtailment were evaluated in a similar way, but were not as efficient. The most inefficient method was futility stopping based on $\hat{\theta}$, which calculates the conditional power assuming $\theta = \hat{\theta}$ and ignores the variability of the estimate.

Most of the work in this chapter went into deriving optimal group sequential designs with non-binding futility boundaries. The aim of this exercise was to enable efficiency comparisons with error spending designs and stochastic curtailment designs that satisfy the same error probability constraints. The optimisation method was outlined in Section 4.3.3, while further details are provided in Section 4.6. Deriving the optimal designs was an interesting challenge, as it was not straightforward to use dynamic programming in the way it was used in Chapter 2. We have nevertheless showed that it is possible to use some of these ideas and extend the dynamic programming method, to solve a more complex problem than is normally possible.

The most important application of the optimal designs is perhaps not to use them in clinical trials, but rather to benchmark other designs such as error spending designs and stochastic curtailment designs. Error spending designs are convenient to use, as they can handle unpredictable group sizes and information levels. We have found that non-binding error spending designs are close to optimal, and therefore recommend their use if non-binding futility boundaries are required. These findings concur with those of Barber and Jennison (2002), for group sequential tests with binding futility boundaries. Stopping for futility if the predictive power is below a certain threshold seems to be another efficient alternative. In order to protect the type I error for unpredictable information sequences, we would recommend to derive the upper boundary from an α spending function. The lower boundary could then be based on either predictive power or a β spending function.

4.6 Derivation of optimal group sequential designs with non-binding futility boundaries

4.6.1 Introduction

We consider a one-sided group sequential test with cumulative information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$. We will show how to derive a group sequential design with a non-binding

futility boundary, that minimises $\tilde{F}$ subject to the desired error probability constraints

$$P_{\theta=0}(\text{Cross upper boundary ignoring lower boundary}) = \alpha \quad (4.16)$$

$$P_{\theta=\delta}(\text{Cross lower boundary before upper boundary}) = \beta. \quad (4.17)$$

To this end, we first consider a Bayes decision problem where we place a three point prior distribution on θ, with probability $1/3$ at $\theta_1 = 0$, $\theta_2 = \delta/2$ and $\theta_3 = \delta$. In our model, θ is generated from this prior distribution, while the sequence of statistics $Z_1, \ldots, Z_K$ follows the usual joint canonical distribution for θ given $\mathcal{I}_1, \ldots, \mathcal{I}_K$, defined in equation (1.1). For $Z_k = z_k$, costs k_{θ_1} and k_{θ_3} are charged as described below, where we assume that $a_k < b_k$, for $k = 1, \ldots, K-1$.

- If a_k and b_k are set so that $z_k \leq a_k$, we say that z_k is in the futility zone. The cost charged is then k_{θ_3} if $\theta = \theta_3$ and neither boundary has been crossed before analysis k, and zero otherwise.
- If a_k and b_k are set so that $a_k < z_k < b_k$, we say that z_k is in the continuation zone.
- If a_k and b_k are set so that $z_k \geq b_k$, we say that z_k is in the efficacy zone. The cost charged is then k_{θ_1} if $\theta = \theta_1$ and the upper boundary has not been crossed before analysis k, and zero otherwise.

We define $C_k = (a_k, b_k)$ and $B_k = (-\infty, b_k)$, for $k = 1, \ldots, K-1$. With a cost of sampling $c_\theta = 1$ per unit information at each value of θ, the total expected cost can be written as

$$\begin{aligned}
&\sum_{i=1}^{3}\sum_{j=1}^{K} \pi_{\theta_i} c_{\theta_i} P_{\theta_i}\{Z_1 \in C_1, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j \text{ or } Z_j \geq b_j\}\mathcal{I}_j \\
+ \;& \pi_{\theta_1} k_{\theta_1} P_{\theta_1}\{Z_j \geq b_j \text{ for some } j = 1, \ldots, K\} \\
+ \;& \pi_{\theta_3} k_{\theta_3} P_{\theta_3}\{Z_1 \in C_1, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j \text{ for some } j = 1, \ldots, K\} \\
\\
= \;& \tilde{F} + \pi_{\theta_1} k_{\theta_1} P_{\theta_1}\{Z_j \geq b_j \text{ for some } j = 1, \ldots, K\} \\
+ \;& \pi_{\theta_3} k_{\theta_3} P_{\theta_3}\{Z_1 \in C_1, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j \text{ for some } j = 1, \ldots, K\} \quad (4.18)
\end{aligned}$$

We will solve the problem of finding the critical values $a_1, b_1, \ldots, a_K, b_K$ that minimise (4.18) through the following steps:

1. Finding a_k and b_k, the critical values at analysis k that minimise (4.18), given critical values $a_1, b_1, \ldots, a_{k-1}, b_{k-1}, a_{k+1}, b_{k+1}, \ldots, a_K, b_K$ and fixed costs k_{θ_1} and k_{θ_3}.

2. Using the method in the previous step, combined with an iterative approach, to find the critical values $a_1, b_1, \ldots, a_K, b_K$ that minimise (4.18), given fixed costs k_{θ_1} and k_{θ_3}.

3. Performing further sensitivity analyses to make sure that the solution that we have found is optimal.

Suppose that we are able to solve the unconstrained decision problem, by finding the critical values $a_1, b_1, \ldots, b_K$ that minimise (4.18), for fixed costs k_{θ_1} and k_{θ_3}. It then remains to perform a numerical search for the costs that give a solution that satisfies the error probability constraints in (4.16) and (4.17). The standard Lagrangian argument implies that this decision rule minimises $\tilde{F}$ among all rules satisfying the same constraints.

4.6.2 Optimising a_k and b_k given critical values at all other analyses

We first consider the problem of finding the critical values a_k, b_k, that solve the unconstrained decision problem defined in the previous section, given boundary points $a_1, b_1, \ldots, a_{k-1}, b_{k-1}, a_{k+1}, b_{k+1}, \ldots, a_K, b_K$ and fixed costs k_{θ_1} and k_{θ_3}. Equation (4.18) can be written as

$$
\begin{aligned}
& \pi_{\theta_1} k_{\theta_1} \Bigg\{ \sum_{j=1}^{k-1} P_{\theta_1}\left\{Z_1 < b_1, \ldots, Z_{j-1} < b_{j-1}, Z_j \geq b_j\right\} \\
+ \;& P_{\theta_1}\left\{Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k \geq b_k\right\} \\
+ \;& \sum_{j=k+1}^{K} P_{\theta_1}\left\{Z_1 < b_1, \ldots, Z_{j-1} < b_{j-1}, Z_j \geq b_j\right\} \Bigg\} \\
+ \;& \pi_{\theta_3} k_{\theta_3} \Bigg\{ \sum_{j=1}^{k-1} P_{\theta_3}\left\{Z_1 \in C_1, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j\right\} \\
+ \;& P_{\theta_3}\left\{Z_1 \in C_1, \ldots, Z_{k-1} \in C_{k-1}, Z_k \leq a_k\right\} \\
+ \;& \sum_{j=k+1}^{K} P_{\theta_3}\left\{Z_1 \in C_1, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j\right\} \Bigg\} \\
+ \;& \sum_{i=1}^{3} \sum_{j=1}^{k-1} \pi_{\theta_i} c_{\theta_i} P_{\theta_i}\{Z_1 \in C_1, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j \text{ or } Z_j \geq b_j\} \mathcal{I}_j \\
+ \;& \sum_{i=1}^{3} \pi_{\theta_i} c_{\theta_i} P_{\theta_i}\{Z_1 \in C_1, \ldots, Z_{k-1} \in C_{k-1}\} \mathcal{I}_k \\
+ \;& \sum_{i=1}^{3} \sum_{j=k}^{K} \pi_{\theta_i} c_{\theta_i} P_{\theta_i}\{Z_1 \in C_1, \ldots, Z_j \in C_j, Z_{j+1} \leq a_{j+1} \text{ or } Z_{j+1} \geq b_{j+1}\} (\mathcal{I}_{j+1} - \mathcal{I}_k).
\end{aligned}
\tag{4.19}
$$

We have divided the terms in (4.19) into three different categories:

1. Terms that do not depend on the critical values at analysis k.
2. Terms that depend on critical values at analysis k, as well as on critical values at previous interim analyses.
3. Terms that are affected by the critical values at analysis k, but are also affected by critical values at previous and later interim analysis.

We shall be making use of the fact that terms in the first category, which later in the derivation will be denoted T_1, T_2 and T_3, do not have to be taken into account when searching for the optimal critical values at analysis k.

Let

$$f_{Z_1,\ldots,Z_k|\theta_i}(z_1,\ldots,z_k|\theta_i)$$

denote the joint pdf of $Z_1,\ldots,Z_k$ given θ_i and let

$$f_{Z_1,\ldots,Z_{k-1}|Z_k,\theta_i}(z_1,\ldots,z_{k-1}|z_k,\theta_i)$$

denote the joint conditional pdf given $Z_k = z_k$ and θ_i. We also define the product sets

$$\boldsymbol{B_k} = B_1 \times B_2 \times \ldots \times B_k$$

and

$$\boldsymbol{C_k} = C_1 \times C_2 \times \ldots \times C_k.$$

We now introduce the notation

$$\begin{aligned}
\int_{\boldsymbol{B_k}} \boldsymbol{dz_k} h_{\theta_i,\boldsymbol{k}}(\boldsymbol{z_k}) &= \int_{B_1} \ldots \int_{B_k} dz_1,\ldots,dz_k f_{Z_1,\ldots,Z_k|\theta_i}(z_1,\ldots,z_k|\theta_i) && (4.20)\\
f_{\theta_i,k}(z_k) &= f_{Z_k|\theta_i}(z_k|\theta_i) && (4.21)\\
f_k(z_k) &= \sum_{i=1}^{3} \pi_{\theta_i} f_{\theta_i,k}(z_k) && (4.22)\\
h_{\boldsymbol{k-1}|k}(\boldsymbol{z_{k-1}}|z_k) &= h_{\theta_i,\boldsymbol{k-1}|k}(\boldsymbol{z_{k-1}}|z_k)\\
&= f_{Z_1,\ldots,Z_{k-1}|Z_k,\theta_i}(z_1,\ldots,z_{k-1}|z_k,\theta_i). && (4.23)
\end{aligned}$$

In (4.23) we have used that since we condition on $Z_k = z_k$, which is a sufficient statistic for θ, the expression must be independent of θ. By using a similar convention as in previous expressions, as well as the fact that we know from (1.1) that the sequence of

test statistics $Z_1, \ldots, Z_K$ has the Markov property, we can write

$$\begin{aligned} h_{\theta_i,k|\boldsymbol{k-1}}(z_k|\boldsymbol{z_{k-1}}) &= f_{Z_k|\boldsymbol{Z_{k-1}},\theta_i}(z_k|\boldsymbol{z_{k-1}},\theta_i) \\ &= f_{Z_k|Z_{k-1},\theta_i}(z_k|z_{k-1},\theta_i) \\ &= h_{\theta_i,k|k-1}(z_k|z_{k-1}). \end{aligned} \tag{4.24}$$

We can now use the notation that has been introduced to write (4.19) as

$$\begin{aligned} & T_1 + k_{\theta_1}\pi_{\theta_1} \int_{\boldsymbol{B_{k-1}}} \boldsymbol{dz_{k-1}} h_{\theta_1,k-1}(\boldsymbol{z_{k-1}}) \int_{-\infty}^{\infty} dz_k h_{\theta_1,k|k-1}(z_k|z_{k-1}) \\ \times \quad & \Big[I(z_k \geq b_k) + (I(z_k < a_k) + I(a_k < z_k < b_k)) \\ \times \quad & \mathrm{P}_{\theta_1}(Z_j \geq b_j \text{ for some } j = k+1, \ldots, K | Z_k = z_k) \Big] \\ + \quad & T_2 + k_{\theta_3}\pi_{\theta_3} \int_{\boldsymbol{C_{k-1}}} \boldsymbol{dz_{k-1}} h_{\theta_3,k-1}(\boldsymbol{z_{k-1}}) \int_{-\infty}^{\infty} dz_k h_{\theta_3,k|k-1}(z_k|z_{k-1}) \\ \times \quad & \Big[I(z_k \leq a_k) + I(a_k < Z_k < b_k) \\ \times \quad & \mathrm{P}_{\theta_3}(Z_{k+1} \in C_{k+1}, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j, \text{ for some } j = k+1, \ldots, K | Z_k = z_k) \Big] \\ + \quad & T_3 + \sum_{i=1}^{3} \pi_{\theta_i} \int_{\boldsymbol{C_{k-1}}} \boldsymbol{dz_{k-1}} h_{\theta_i,\boldsymbol{k-1}}(\boldsymbol{z_{k-1}}) \int_{-\infty}^{\infty} dz_k h_{\theta_i,k|k-1}(z_k|z_{k-1}) \\ \times \quad & I(a_k < Z_k < b_k) E_{\theta_i}(\mathcal{I}_{final} - \mathcal{I}_k | Z_k = z_k), \end{aligned} \tag{4.25}$$

where $\mathcal{I}_{final}$ is the information on termination of the group sequential test, assuming that the futility boundary is always applied, and the terms

$$\begin{aligned} T_1 &= k_{\theta_1}\pi_{\theta_1} \sum_{j=1}^{k-1} P_{\theta_1}\{Z_1 < b_1, \ldots, Z_{j-1} < b_{j-1}, Z_j \geq b_j\} \\ T_2 &= k_{\theta_3}\pi_{\theta_3} \sum_{j=1}^{k-1} P_{\theta_3}\{a_1 < Z_1 < b_1, \ldots, a_{j-1} < Z_{j-1} < b_{j-1}, Z_j \leq a_j\} \\ T_3 &= \sum_{i=1}^{3} \sum_{j=1}^{k-1} \pi_{\theta_i} c_{\theta_i} P_{\theta_i}\{Z_1 \in C_1, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j \text{ or } Z_j \geq b_j\} \mathcal{I}_j \\ &\quad + \sum_{i=1}^{3} \pi_{\theta_i} c_{\theta_i} P_{\theta_i}\{Z_1 \in C_1, \ldots, Z_{k-1} \in C_{k-1}\} \mathcal{I}_k \end{aligned}$$

do not depend on the critical values at analysis k.

We shall now derive some properties that can be used to further simplify (4.25). We can combine (4.23) and (4.24) to obtain

$$\begin{aligned} h_{\theta,k|k-1}(z_k|z_{k-1})h_{\theta,\boldsymbol{k-1}}(\boldsymbol{z_{k-1}}) &= h_{\theta,k|\boldsymbol{k-1}}(z_k|\boldsymbol{z_{k-1}})h_{\theta,\boldsymbol{k-1}}(\boldsymbol{z_{k-1}}) \\ &= f_{\theta,k}(z_k)h_{\theta,\boldsymbol{k-1}|k}(\boldsymbol{z_{k-1}}|z_k) \\ &= f_{\theta,k}(z_k)h_{\boldsymbol{k-1}|k}(\boldsymbol{z_{k-1}}|z_k). \end{aligned} \tag{4.26}$$

We will now use the conditional probabilities

$$p_{kc}(z_k) = P(Z_1 \in C_1, \ldots, Z_{k-1} \in C_{k-1}|Z_k = z_k) \tag{4.27}$$

$$p_{kb}(z_k) = P(Z_1 \in B_1, \ldots, Z_{k-1} \in B_{k-1}|Z_k = z_k), \tag{4.28}$$

which for the same reason as $h_{\boldsymbol{k-1}|k}(\boldsymbol{z_{k-1}}|z_k)$ in (4.23) are independent of θ. Finally we will use that

$$\pi_{\theta_i} f_{\theta_i,k}(z_k) = \frac{\pi_{\theta_i} f_{\theta_i,k}(z_k) f_k(z_k)}{f_k(z_k)} = \pi^{(k)}(\theta_i|z_k) f_k(z_k), \tag{4.29}$$

where $\pi^{(k)}(\theta_i|z_k)$ denotes the density of the posterior distribution of θ_i given $Z_k = z_k$ and consequently, the last equality follows from Bayes theorem. By combining (4.26), (4.27) and (4.29), it follows that for any function $g(z_k)$,

$$\begin{aligned} & \pi_{\theta_1} \int_{\boldsymbol{B_{k-1}}} \boldsymbol{dz_{k-1}} h_{\theta_1,\boldsymbol{k-1}}(\boldsymbol{z_{k-1}}) \int_{-\infty}^{\infty} dz_k h_{\theta_1,k|k-1}(z_k|z_{k-1}) g(z_k) \\ = \; & \pi_{\theta_1} \int_{-\infty}^{\infty} dz_k f_{\theta_1,k}(z_k) g(z_k) \int_{\boldsymbol{B_{k-1}}} \boldsymbol{dz_{k-1}} h_{\boldsymbol{k-1}|k}(\boldsymbol{z_{k-1}}|z_k) \\ = \; & \int_{-\infty}^{\infty} dz_k f_k(z_k) \pi^{(k)}(\theta_1|z_k) g(z_k) \int_{\boldsymbol{B_{k-1}}} \boldsymbol{dz_{k-1}} h_{\boldsymbol{k-1}|k}(\boldsymbol{z_{k-1}}|z_k) \\ = \; & \int_{-\infty}^{\infty} dz_k f_k(z_k) \pi^{(k)}(\theta_1|z_k) g(z_k) P(Z_1 \in B_1, \ldots, Z_{k-1} \in B_{k-1}|Z_k = z_k) \\ = \; & \int_{-\infty}^{\infty} dz_k f_k(z_k) \pi^{(k)}(\theta_1|z_k) p_{kb}(z_k) g(z_k) \end{aligned} \tag{4.30}$$

and in the same way, for any function $g(z_k)$, that

$$\begin{aligned} & \pi_{\theta_3} \int_{\boldsymbol{C_{k-1}}} \boldsymbol{dz_{k-1}} h_{\theta_3,\boldsymbol{k-1}}(\boldsymbol{z_{k-1}}) \int_{-\infty}^{\infty} dz_k f_{\theta_3,k}(z_k|z_{k-1}) g(z_k) \\ = \; & \int_{-\infty}^{\infty} dz_k f_k(z_k) \pi^{(k)}(\theta_3|z_k) p_{kc}(z_k) g(z_k). \end{aligned} \tag{4.31}$$

Finally, we can now insert (4.30) and (4.31) into equation (4.25) to obtain

$$\begin{aligned}
& T_1 + T_2 + T_3 + \int_{-\infty}^{\infty} \Big\{ dz_k f_k(z_k) \Big\{ p_{kb}(z_k) k_{\theta_1} \pi^{(k)}(\theta_1|z_k) \Big[I(z_k \geq b_k \\
+ \quad & (I(z_k \leq a_k) + I(a_k < z_k < b_k)) \mathrm{P}_{\theta_1}(Z_j \geq b_j \text{ for some } j = k+1, \ldots, K | Z_k = z_k) \Big] \\
+ \quad & p_{kc}(z_k) k_{\theta_3} \pi^{(k)}(\theta_3|z_k) \Big[I(z_k \leq a_k) + I(a_k < z_k < b_k) \\
\times \quad & \mathrm{P}_{\theta_3}(Z_{k+1} \in C_{k+1}, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j, \text{ for some } j = k+1, \ldots, K | Z_k = z_k) \Big] \\
+ \quad & I(a_k < Z_k < b_k) p_{kc}(z_k) \sum_{i=1}^{3} \pi^{(k)}(\theta_i|z_k) E_{\theta_i}(\mathcal{I}_{final} - \mathcal{I}_k) | Z_k = z_k) \Big\} \Big\}. \qquad (4.32)
\end{aligned}$$

We now consider $Z_k = z_k$ and in which zone z_k should be put, in order for (4.32) to be minimised. It follows from the definition of T_1, T_2 and T_3 that these terms are not affected by the critical values at analysis k, so they can be disregarded. By comparing the terms multiplying the indicator functions $I(z_k \leq a_k)$, $I(z_k \geq b_k)$ and $I(a_k < z_k < b_k)$, we find the following:
If we put z_k in the futility zone, the expected additional cost incurred is proportional to

$$\begin{aligned}
& p_{kb}(z_k) k_{\theta_1} \pi^{(k)}(\theta_1|z_k) \mathrm{P}_{\theta_1}(Z_j \geq b_j \text{ for some } j = k+1, \ldots, K | Z_k = z_k) \\
+ \quad & p_{kc}(z_k) k_{\theta_3} \pi^{(k)}(\theta_3|z_k). \qquad (4.33)
\end{aligned}$$

If we put z_k in the continuation zone, the expected additional cost incurred is proportional to

$$\begin{aligned}
& p_{kc}(z_k) k_{\theta_3} \pi^{(k)}(\theta_3|z_k) \\
\times \quad & \mathrm{P}_{\theta_3}(Z_{k+1} \in C_{k+1}, \ldots, Z_{j-1} \in C_{j-1}, Z_j \leq a_j, \text{ for some } j = k+1, \ldots, K | Z_k = z_k) \\
+ \quad & p_{kb}(z_k) k_{\theta_1} \pi^{(k)}(\theta_1|z_k) \mathrm{P}_{\theta_1}(Z_j \geq b_j \text{ for some } j = k+1, \ldots, K | Z_k = z_k) \\
+ \quad & p_{kc}(z_k) \sum_{i=1}^{3} \pi^{(k)}(\theta_i|z_k) E_{\theta_i}(\mathcal{I}_{final} - \mathcal{I}_k) | Z_k = z_k). \qquad (4.34)
\end{aligned}$$

If we put z_k in the efficacy zone, the expected additional cost incurred is proportional to

$$p_{kb}(z_k) k_{\theta_1} \pi^{(k)}(\theta_1|z_k). \qquad (4.35)$$

It is intuitive to expect that the optimal solution is for z_k to be in the futility zone for the lowest values, the continuation zone for intermediate values, and in the efficacy zone for high values. In this case these definitions agree with the form of boundary

specified and the same a_k and b_k values are ideal for all z_k. In fact, this is what we have found for all the cases that we have considered. The critical value a_k can then be found as the root of z_k where (4.33) equals (4.34) and b_k as the root of z_k where (4.34) equals (4.35). Further details for how this can be achieved are discussed in Section 4.6.4.

4.6.3 Finding the optimal critical values at all analyses

In Section 4.6.2 we showed how to find the critical values a_k and b_k by assuming the critical values at all other analyses to be known. Mathematically, the knowledge gained in Section 4.6.2 can be expressed in terms of functions ϕ_k and ψ_k. We have

$$a_k = \phi_k(a_1, b_1, \ldots, a_{k-1}, b_{k-1}, a_{k+1}, b_{k+1}, \ldots, a_K, b_K)$$

and

$$b_k = \psi_k(a_1, b_1, \ldots, a_{k-1}, b_{k-1}, a_{k+1}, b_{k+1}, \ldots, a_K, b_K)$$

where a_K is added for notational convenience and is equal to b_K. By assuming starting values

$$\{a^{(0)}, b^{(0)}\}_1^{K-1} = a_1^{(0)}, b_1^{(0)}, \ldots, a_{K-1}^{(0)}, b_{K-1}^{(0)},$$

we can find the optimal solution $b_K^{(1)}$ at analysis K, for a given set of starting values $\{a^{(0)}, b^{(0)}\}_1^{K-1}$. The optimal decision rule at analysis $K-1$ can thereafter be found using $\{a, b\}_0^{K-2}$, as well as the solution $b_K^{(1)}$ that was just derived. For general k, the critical values are found using $\{a^{(0)}, b^{(0)}\}_1^{k-1}$ and $\{a^{(1)}, b^{(1)}\}_{k+1}^{K}$. This procedure can be used iteratively to create a sequence of solutions

$$\{a^{(1)}, b^{(1)}\}_1^K, \ldots, \{a^{(m)}, b^{(m)}\}_1^K.$$

For each iteration, we can improve on the previous solution. At iteration i we have

$$a_k^{(i)} = \phi(\{a^{(i-1)}, b^{(i-1)}\}_1^{k-1}, \{a^{(i)}, b^{(i)}\}_{k+1}^{K})$$

and

$$b_k^{(i)} = \psi(\{a^{(i-1)}, b^{(i-1)}\}_1^{k-1}, \{a^{(i)}, b^{(i)}\}_{k+1}^{K}).$$

When

$$\sum_{k=1}^{K} \left((a_k^{(i+1)} - a_k^{(i)})^2 + (b_k^{(i+1)} - b_k^{(i)})^2 \right) < \epsilon,$$

where $\epsilon \approx 10^{-6}$, we consider the method to have converged. This typically happens after fewer than ten iterations. In principle there is a risk that the method has not converged to a global minimum, but we have not found any such problems. In Section 4.6.4 we will discuss sensitivity checks that can be performed to assess further

whether the solution could be improved upon.

To solve the more general problem of finding the optimal solution $a_1, b_1, \ldots, a_K, b_K$, we had to calculate the conditional probabilities in equations (4.27) and (4.28). These calculations require knowledge of critical values at earlier interim analyses that are unknown, which is why we need to assume starting values and use the iterative procedure. All of this is done assuming a set of starting points $\{a^{(0)}, b^{(0)}\}_1^{K-1}$. We know from Barber and Jennison (2002) how to find the optimal boundary when the futility boundary is binding. When solving our problem with a non-binding futility boundary, we have found it convenient to use the critical values of the solution to the corresponding binding problem as starting values.

4.6.4 Implementation and sensitivity checks

When implementing this method computationally, it is useful to define and store the expected cost of setting z_k in the futility zone, efficacy zone or continuation zone. It is clear from (4.32) that the conditional probabilities, $p_{kc}(z_k)$ in (4.27) and $p_{kb}(z_k)$ in (4.28), are needed to derive the optimal decision rule at each stage. As knowledge about the critical values $a_1, b_1, \ldots, a_{k-1}, b_{k-1}$ at previous analyses are needed to calculate $p_{kc}(z_k)$ and $p_{kb}(z_k)$, it is not obvious that it would be beneficial to start by finding the optimal decision at analysis K. One could in principle start by deriving the critical values at any analysis k, conditional on some assumed starting values for the critical values at other analyses. We have nevertheless found it convenient to start at the final analysis and work backwards.

We calculate $p_{kc}(z_k)$ and $p_{kb}(z_k)$ through a forward calculation, storing the sub-densities at each analysis on a grid of values for z_k. We are typically interested in the ratio $p_{kc}(z_k)/p_{kb}(z_k)$, which does not depend on θ. So this calculation can be done for any value of θ, $\theta = 0$ say. When it is necessary to calculate $p_{kc}(z_k)/p_{kb}(z_k)$ for a specific $Z_k = z_k$ that was not included in the grid, we can make use of the sub-densities that have been stored and only have to calculate the transition from analysis $k-1$ to analysis k.

We have described a method to find the optimal decision rule. It has been found, in all cases, to satisfy the natural property that for $k = 1, \ldots, K-1$, there is an interval $(-\infty, a_k]$ where it is optimal to put z_k in the futility zone, followed by an interval (a_k, b_k) where it is optimal to put z_k in the continuation zone, and that finally there is an interval $[b_k, \infty)$ where it is optimal to put z_k in the efficacy zone. This appears to be a reasonable assumption, but one might ask if another pattern could be possible. Numerical checks have been built into our computer program to assess whether this is the case, but no examples that violate these assumptions have been found.

We start by assuming starting values for $\{a^{(0)}, b^{(0)}\}_1^{K-1}$, so that the root b_K at

the final analysis can be calculated. The other critical values can thereafter be calculated recursively, working backwards from analysis $K-1$: at analysis k we use the stage $k+1$ critical values, the functions $p_{kc}(z_k)$ and $p_{kb}(z_k)$ calculated for the starting values as well as the expected costs of setting z_{k+1} in the futility zone, continuation zone or efficacy zone, previously calculated on a grid of z_{k+1} values. To cope with discontinuities, the points a_{k+1} and b_{k+1} are added to the grid.

The integrals in (4.32) can be calculated by numerical integration using Simpson's rule. At each analysis k, the roots where equation (4.33) equals equation (4.34) and where equation (4.34) equals equation (4.35), are found by a numerical search, and these two roots define the stopping boundaries. When searching for a root, b_k say, we make use of the grid for z_k that is available from the numerical integration. The root b_k is found by searching for when the expression

$$\begin{aligned}
& p_{kc}(z_k)k_{\theta_3}\pi^{(k)}(\theta_3|z_k) \\
\times\ & \mathrm{P}_{\theta_3}(Z_{k+1}\in C_{k+1},\ldots,Z_{j-1}\in C_{j-1},Z_j\le a_j,\ \text{for some } j=k+1,\ldots,K|Z_k=z_k) \\
+\ & p_{kb}(z_k)k_{\theta_1}\pi^{(k)}(\theta_1|z_k)\mathrm{P}_{\theta_1}(Z_j\ge b_j \text{ for some } j=k+1,\ldots,K|Z_k=z_k) \\
+\ & p_{kc}(z_k)\sum_{i=1}^{3}\pi^{(k)}(\theta_i|z_k)E_{\theta_i}(\mathcal{I}_{final}-\mathcal{I}_k)|Z_k=z_k) \\
-\ & p_{kb}(z_k)k_{\theta_1}\pi^{(k)}(\theta_1|z_k) \qquad (4.36)
\end{aligned}$$

changes value between two grid points. We then perform a numerical search for the root where (4.36) equals zero. Other grid points are found in a similar way.

We are relying on starting values and an iterative procedure to find our optimal solution. It is then important to perform sensitivity checks and assess the robustness of our results. One sensitivity check that has been performed is to solve the unconstrained decision problem by performing a direct numerical search over the set of boundary points, using the simplex method of Nelder and Mead (1965). Our experience is that for reasonable starting values, the direct search method typically finds the same solution as our method. Another sensitivity check has been performed by using the solution derived in Section 4.6.2 as starting value and applying the direct numerical search to see if an improvement is possible. In no case has the direct search method provided a solution that improves on our optimisation method.

We have now outlined the general principles for finding the critical values $a_1, b_1, \ldots, a_K, b_K$ that solve the unconstrained Bayes decision problem defined in (4.18), given fixed costs k_{θ_1} and k_{θ_3}. It remains to perform a search for the costs k_{θ_1} and k_{θ_3} that give the desired error probabilities α and β. The standard Lagrangian argument then implies that our optimal solution minimises $\tilde{F}$ among all rules satisfying the same constraints.

CHAPTER 5

Joint planning of phase II and phase III

5.1 Introduction

The problems considered in previous chapters all deal with the design of separate trials. In this chapter we move beyond the individual trial and consider the joint planning of phase II and phase III trials. We consider how to approach the design of both stages, with a focus on what one stage contributes to the design of the other. We define a utility function that includes a reward if regulatory approval is achieved after phase III, and the costs of the phase II and phase III trials. The objective is to design one phase II trial and one phase III trial, in a way that maximises the expected utility. The phase II trial is assumed to be a fixed sample trial, while the phase III trial will be either a fixed sample trial or group sequential. We shall be focusing on two-arm phase II trials comparing one experimental drug and a control, thus not addressing the issue of dose-finding.

We assume that different endpoints are used in phase II and phase III. The relationship between the means of the phase II and phase III endpoints is modelled within a Bayesian framework. As phase III trials are typically more expensive and complex to run, it is natural to consider the role of phase II as a possibility to choose which drugs are really worth this investment, and which should be discontinued. This is likely to be of particular value if the information gathered in phase II can be obtained more cheaply and in a shorter time frame, while still being able to predict the phase III outcome with adequate accuracy.

While the design of phase II and phase III clinical trials has attracted a lot of interest in the statistical literature, most publications primarily focus on one of the phases. Schoenfeld (1980) provides statistical considerations for phase II trials, with a particular focus on applications in oncology. He emphasises that the decision rule for progress to phase III should be specified in the study protocol and taken into

account when deciding the phase II sample size. Whitehead (1985) discusses the role of the phase II trial in a clinical development programme. He considers the situation where several treatments are tested in phase II, and one of these is selected to progress to phase III. Taking the phase III sample size as fixed, Stallard (1998) uses explicit cost and gain functions to derive optimal group sequential phase II designs for binary outcomes. This method is generalised by Stallard et al. (1999), to consider both efficacy and toxicity. A good overview about different approaches to the design of phase II trials is given by Stallard et al. (2001).

Other publications take the design of phase II as given, but discuss how to use the results in phase II to guide the design of phase III. Wang et al. (2006) consider the problem of choosing sample size for phase III based on phase II data, when the same endpoint is used in the two phases. They note that simply assuming that the treatment effect in phase III is equal to the point estimate of the treatment effect in phase II may be too optimistic. Instead, it is advocated to take a more conservative approach. One possibility that is mentioned by Wang et al. (2006) is to use the point estimate of the treatment effect minus one standard error, as the treatment effect in the sample size calculation for the new trial. Pezeshk et al. (2009) focus on how to design phase III using a mixed Bayesian and frequentist approach. They consider the design of a single phase III trial and assume that the treatment effect has a certain prior distribution. They do not discuss the design of phase II but the prior distribution for the phase III treatment effect can be thought of as being derived after incorporating the phase II results. A frequentist hypothesis test is performed at the end of phase III.

A more holistic approach is taken in so-called seamless phase II/III trials, where data from phase II and phase III are combined in the final analysis. Different approaches for how to design such trials have been proposed by Bauer and Kieser (1999) and Stallard and Todd (2003), among others. The papers by Bretz et al. (2006) and Schmidli et al. (2006) describe a method for data combination that is based on the p-value combination tests of Bauer and Köhne (1994). The latter of the two papers also includes practical considerations for how the method can be applied in practice. The duration of phase II trials is often too short to be able to observe the long-term clinical endpoint at the end of the trial. Jenkins et al. (2011) give an example of how long-term follow-up data for the phase II patients can be combined with the phase III data, despite not being available at the end of phase II. They ensure control of the type I error rate by using the combination rule of Bauer and Köhne (1994), and letting the long-term follow-up of the phase II subjects contribute to the stage 1 p-value. A different approach is used by Stallard (2010), who suggests controlling the type I error by adjusting the group sequential boundary in the confirmatory phase.

We shall however focus on the more conventional situation, where data from different phases are not combined in the final analysis. It is then only the phase III

data that contribute to the final hypothesis test in phase III. Whitehead (1986) and Antonijevic et al. (2010), among others, consider the joint planning of phase II and phase III in this situation. Whitehead (1986) provides an integrated approach to phases II and III in the situation where several treatments are available in phase II. Several choices of phase III sample size are considered, but the phase III sample size is not optimised. Antonijevic et al. (2010) discuss how dose selection strategies in phase II affect the probability of success in phase III. While the phase III sample size is held fixed, different choices for phase II sample size are compared. Our model differs from Whitehead (1986) and Antonijevic et al. (2010) in that the sample sizes for both phase II and phase III are optimised. Furthermore, we consider the possibility of running a group sequential design in phase III, and investigate how this impacts the phase II sample size.

In this chapter we shall consider the joint planning of one phase II trial and one phase III trial, comparing a single treatment to a control in both trials. We seek to optimise the phase II and phase III sample sizes, n_2 and n_3. In two-arm phase II clinical trials, comparing an experimental drug to a control, the most important decision is arguably whether to progress the drug under investigation to phase III. In addition, the phase II results may help to guide the design of phase III. Hence, we shall also consider how to find an optimal decision rule for the go/no go decision, and how to choose the phase III sample size, based on phase II data. We will start by assuming fixed sample trials in both phase II and phase III, but later move on to consider the possibility of a group sequential phase III trial.

In Section 5.2, we describe the Bayesian framework that is used in this chapter. Section 5.2 is also where the expected utility that we seek to maximise is defined. Thereafter, the optimisation of phase II, for a given phase III design, is studied in Section 5.3. In Section 5.4 it is shown how the concept of optimally purchasing information can be applied, for example to guide a decision about which biomarker to use in phase II. In Section 5.5 we discuss the joint optimisation of both phases, including the case when phase III is group sequential. A numerical example is provided in Section 5.6, while discussion and conclusions are provided in Section 5.7. Finally, further details of how the model is derived and implemented numerically are given in Section 5.8.

5.2 Model

5.2.1 Introduction

As we are interested in designing a phase II trial in which the endpoint is different from that measured in phase III, we need a model for how the phase II and phase III outcomes relate to each other. Hence, this section starts with a short discussion about

different types of endpoints and their use in drug development. Thereafter, we present the definition of our model and discuss what is required for regulatory approval. Finally, the costs of conducting the trials and the gain from obtaining regulatory approval are introduced and brought together into a formula. This formula defines the expected utility that we seek to optimise.

5.2.2 Biomarkers, surrogate endpoints and clinical endpoints

Lesko and Atkinson (2001) give a comprehensive overview of biomarkers and surrogate endpoints. They define a biomarker as a physical sign or laboratory measurement that occurs in association with a pathological process and that has putative diagnostic and/or prognostic utility. A surrogate endpoint is, according to the same authors, a biomarker that is intended to serve as a substitute for a clinically meaningful endpoint and is expected to predict the effect of a therapeutic intervention. Finally, a clinical endpoint is defined as a clinically meaningful measure of how a patient feels, functions or survives.

According to the definitions of Lesko and Atkinson (2001), a biomarker is a surrogate endpoint only if certain criteria are met. These criteria vary between different authors, but the rather restrictive definition of Prentice (1989) is one that is frequently cited. Prentice (1989) requires that the endpoint is correlated with the true clinical outcome in individual subjects. In addition, the surrogate endpoint should capture the net effect of the treatment on the clinical endpoint so that finding a treatment effect, relative to placebo say, on the surrogate endpoint implies a strong likelihood of a treatment effect on the clinical endpoint. Fleming and DeMets (1996) describe situations where the latter condition does not apply and how this has led to disappointing results once the long term effect on the clinical endpoint could be evaluated in clinical trials. The view expressed by ICH (1998) reflects the importance of the issues raised by Prentice (1989) and Fleming and DeMets (1996). ICH (1998) recommends the following criteria for establishing the strength of evidence for an endpoint to act as a surrogate:

- biological plausibility of the relationship,
- the demonstration in epidemiological studies of the prognostic value of the surrogate endpoint,
- evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome.

If a surrogate endpoint is well established, it may no longer be necessary to demonstrate efficacy for the clinical endpoint in confirmatory phase III trials. This is the case for conditions such as hypertension and hypercholesterolemia, where regulatory

authorities have typically accepted the use of reduction in blood pressure or LDL-cholesterol respectively, as surrogate endpoints. When a surrogate endpoint replaces the clinical endpoint as the primary endpoint in phase III, it is sometimes the case that the effect on clinical endpoints such as death, myocardial infarction and stroke are studied in phase IV trials, carried out after regulatory approval.

For conditions where the regulatory authorities approve drugs only if efficacy on the long-term clinical endpoint has been demonstrated, biomarkers still have a role to play. As we have already pointed out, it may be very costly and time-consuming to measure the clinical endpoint in phase II. We will in this chapter study, among other things, when it could be beneficial to study a biomarker in phase II. The option of using the biomarker in phase II may be particularly appealing if the response can be obtained within a short time frame and is less costly to measure.

5.2.3 Model for phase II and phase III

Suppose that the variable X is measured in a phase II clinical trial, with n_2 patients randomised to each of the treatment groups A and B. The observations are conditionally independent given θ_{2B} and θ_{2A} and normally distributed with $X_{Ai} \sim N(\theta_{2A}, \sigma_2^2)$ and $X_{Bi} \sim N(\theta_{2B}, \sigma_2^2)$, where the common variance σ_2^2 is known. The parameter of interest in the phase II trial is $\theta_2 = \theta_{2B} - \theta_{2A}$. Learning about θ_2 can be beneficial if the knowledge gained can be translated into information about the clinical endpoint Y, measured in phase III. In the phase III trial with treatment groups A and B, there are n_3 observations per group that are conditionally independent given θ_{3B} and θ_{3A} and normally distributed according to $Y_{Ai} \sim N(\theta_{3A}, \sigma_3^2)$ and $Y_{Bi} \sim N(\theta_{3B}, \sigma_3^2)$, with known common variance σ_3^2. At the end of phase III, we are interested in testing the null hypothesis $\theta_3 = \theta_{3B} - \theta_{3A} \leq 0$. Hence, it is important to know how θ_2 relates to θ_3, when evaluating different options for how to design phase II. The joint prior distribution of (θ_2, θ_3) is assumed to be multivariate normal according to

$$(\theta_2, \theta_3) \sim N\left((m_1, \mu_1), \begin{pmatrix} t_1^2 & rt_1\tau_1 \\ rt_1\tau_1 & \tau_1^2 \end{pmatrix}\right), \tag{5.1}$$

where m_1 and μ_1 are the prior means and t_1^2 and τ_1^2 the prior variances, of θ_2 and θ_3 respectively. The correlation between θ_2 and θ_3 is denoted r and plays an important role in our model. If (5.1) holds, the conditional distribution

$$(\theta_3|\theta_2) \sim N(C + D\theta_2, \tau_\epsilon^2) \tag{5.2}$$

is obtained, where $C = \mu_1 - rm_1\tau_1/t_1$, $D = r\tau_1/t_1$ and $\tau_\epsilon^2 = (1-r^2)\tau_1^2$. The correlation r thus determines to what degree the uncertainty of θ_3 can be reduced by learning about θ_2.

After the phase II trial, we are interested in updating the prior distribution with the new information obtained from the phase II trial. Let us first define

$$Z_2 \quad = \quad \bar{X}_{n_2}\sqrt{n_2/(2\sigma_2^2)},$$

where $\bar{X}_{n_2} = \bar{X}_B - \bar{X}_A$ is based on n_2 observations per treatment group. We make the modelling assumption that Z_2 is conditionally independent of θ_3 given θ_2, so for the pdf of Z_2 given θ_2 and θ_3 we have

$$f_{Z_2|\theta_2,\theta_3}(z_2|\theta_2,\theta_3) = f_{Z_2|\theta_2}(z_2|\theta_2), \tag{5.3}$$

where

$$Z_2|\theta_2 \sim N(\theta_2\sqrt{n_2/(2\sigma_2^2)}, 1).$$

Once the phase II results are available, we can given $Z_2 = z_2$ calculate the posterior distribution of θ_2. As shown in standard textbooks about Bayesian Statistics, such as the book by Gelman et al. (2004), it follows a normal distribution according to

$$\theta_2|Z_2 = z_2 \sim N(m_2, t_2^2),$$

where the mean m_2 and the variance t_2^2 can be written as

$$m_2 \quad = \quad \frac{m_1/t_1^2 + \bar{x}_{n_2}n_2/(2\sigma_2^2)}{1/t_1^2 + n_2/(2\sigma_2^2)} = \frac{m_1/t_1^2 + z_2\sqrt{n_2/(2\sigma_2^2)}}{1/t_1^2 + n_2/(2\sigma_2^2)}, \tag{5.4}$$

$$t_2^2 \quad = \quad (1/t_1^2 + n_2/(2\sigma_2^2))^{-1}. \tag{5.5}$$

No observations of the phase III endpoint are observed in phase II, but we would still like to update the prior distribution of θ_3 with the results for the phase II endpoint X. If follows immediately from (5.3) that

$$\pi_{\theta_3|\theta_2,Z_2}(\theta_3|\theta_2,z_2) = \pi_{\theta_3|\theta_2}(\theta_3|\theta_2), \tag{5.6}$$

since θ_3 is conditionally independent of Z_2 given θ_2. We show in Section 5.8.1 that (5.6) in combination with (5.2) can be used to find the posterior distribution for θ_3 after phase II. It is normally distributed according to

$$\theta_3|Z_2 = z_2 \quad \sim \quad N(\mu_2, \tau_2^2), \tag{5.7}$$

where the mean μ_2 and the variance τ_2^2 can be written as

$$\mu_2 = \mu_2(n_2, z_2) = C + Dm_2 \tag{5.8}$$
$$\tau_2^2 = \tau_2^2(n_2) = \tau_\epsilon^2 + D^2 t_2^2. \tag{5.9}$$

After phase II, the phase III sample size n_3 can in the most general version of our model be chosen based on the observed value of Z_2 and the phase II sample size n_2. The phase III sample size is then a function of $Z_2 = z_2$ and n_2, so when appropriate we shall write $n_3(z_2, n_2)$ to make this explicitly clear. In phase III, the standardised statistic

$$Z_3 = \bar{Y}_{n_3}\sqrt{n_3/(2\sigma_3^2)}, \tag{5.10}$$

where $\bar{Y}_{n_3} = \bar{Y}_B - \bar{Y}_A$ is based on n_3 observations per group, is compared against a critical value $z_{1-\alpha}$. We make the modelling assumption that Z_3 is conditionally independent of θ_2 given θ_3 and Z_2. Hence, for the probability density function of Z_3 given θ_3, θ_2 and z_2, we have

$$f_{Z_3|\theta_3,\theta_2,Z_2}(z_3|\theta_3,\theta_2,z_2) = f_{Z_3|\theta_3,Z_2}(z_3|\theta_3,z_2), \tag{5.11}$$

where the observed value of Z_2 effects the distribution of Z_3 through n_3. The distribution of Z_3 given θ_3 and $Z_2 = z_2$ is normal according to

$$Z_3|\theta_3, Z_2 = z_2 \sim N(\theta_3\sqrt{n_3(z_2,n_2)/(2\sigma_3^2)}, 1).$$

Since θ_3 denotes the difference in means between the two treatment groups in phase III it is appropriate to plan phase III based on the posterior distribution of θ_3 after phase II, given in equation (5.7).

Figure 5-1 illustrates how t_2^2 and τ_2^2 decrease with increasing phase II sample size n_2, for the case when $r = 0.8$, $\sigma_2^2 = 1$, $t_1^2 = 0.04$ and $\tau_1^2 = 0.04$. By running a large enough phase II trial, the uncertainty about θ_2 can be completely eliminated. The prior variance of θ_3 can however only be reduced from τ_1^2 to a minimum of $(1-r^2)\tau_1^2$, achieved as $n_2 \to \infty$. We see that r^2 appears in the expression for the posterior variance of θ_3 after phase II. In the remainder of this chapter we shall focus on correlations that are positive, i.e. $r \in [0, 1]$. We note however that biomarkers with correlation $-r$ would be just as useful in reducing the posterior variance as biomarkers with correlation of r.

Yin (2002) models the phase II and phase III outcomes in a Bayesian framework that bears similarities with our model. One difference between Yin's model and ours is that we consider the joint optimisation of the sample sizes to be used in phase II and phase III. In Yin's model, the focus is on the sample size of the phase II trial, while the phase III sample size is taken as fixed. Another difference is that in the model of

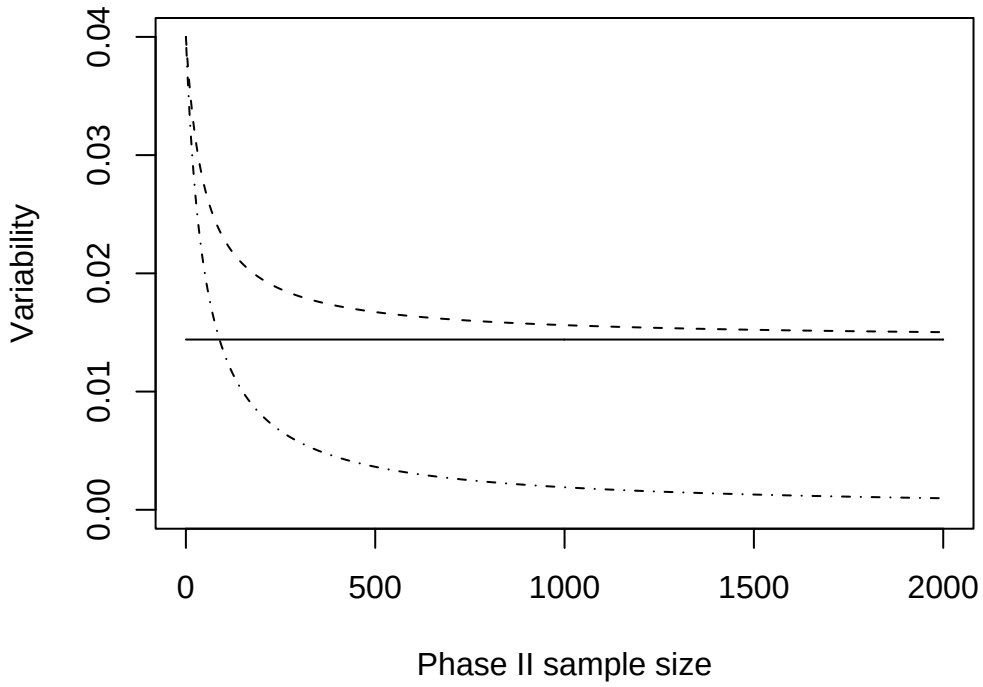

Figure 5-1: Posterior variance of θ_3 (dashed line) and θ_2 (dot-dashed line) after phase II, when $r = 0.8$, $\sigma_2^2 = 1$, $t_1^2 = 0.04$ and $\tau_1^2 = 0.04$. The solid horizontal line shows $\tau_\epsilon^2 = (1 - r^2)\tau_1^2$.

Yin (2002), the long-term phase III outcome depends on the short-term outcome in phase III on a patient level. The uncertainty of the short-term outcome in phase III and its relationship with the short-term outcome in the phase II trial is expressed in a Bayesian model, which is updated after the phase II trial. Inoue et al. (2002) and Todd and Stallard (2005) also consider a change of endpoint between the two stages, in the context of seamless phase II/III designs.

5.2.4 Requirements for regulatory approval

Two independent phase III trials, each with a statistically significant result, typically $p < 0.025$ one-sided, are often required for regulatory approval. This requirement is sometimes referred to as the two trials rule. Senn (1997) and Bauer (2003) have commented on the fact that one phase III trial with a one-sided type I error probability of α^2 would give the same risk of an inefficacious drug being approved, as applying the two trials rule with a one-sided type I error probability of α. Bauer (2003) also considers the increase in false-positive rate if two positive trials out of a total of three or four are

required. Senn (1997) argues that if the two trials rule is rigorously applied and fully taken into account at the planning stage, several aspects of drug development may have to change. Power may have to be set higher for individual trials, while the conduct of group sequential trials would also be affected.

Senn (1997) refers to the approach of focusing on one trial with a lower type I error, rather than the two trials rule, as the pooled trials rule. A solution that will be used in this paper is, instead of requiring two trials that both are significant at the $\alpha = 0.025$ level, to focus on one trial with type I error probability $\alpha = 0.0005 < 0.025^2$. If $\alpha = 0.025^2$ is allowed for a one-sided test, it is reasonable to allow $2 \times 0.025^2 = 0.00125$ for a two-sided test. For a two-sided test, 0.00125 can be suitably rounded downwards to 0.001, which would give us 0.0005 for a one-sided test. In practice there should be very little difference between using a type I error probability of 0.0005 and 0.000625. Considering only one trial makes it easier to account for potential savings with group sequential designs. Achieving a result that is statistically significantly different from control, at the 0.0005 level one-sided, will be the threshold for regulatory approval in our model. We do not make any requirement for safety, even though in practice a positive benefit/risk assessment of the new drug would be required to obtain regulatory approval.

It is important to note that our model of replacing two phase III trials with one phase III trial must be viewed as an approximation. Even though the type I error is adjusted to be the same for the two procedures, there are other differences that are worth mentioning. Consider the case with fixed sample two-arm trials, either one trial with type I error probability α^2 and n_3 patients per group, or two trials, each with type I probability α and $n_3/2$ patients per treatment group. It then follows from the Neyman-Pearson lemma that it must be more efficient to base decisions on the sufficient statistic for the whole sample, rather than calculating separate statistics for the two independent trials and requiring $Z \geq z_{1-\alpha}$ for both of these sub-samples.

Let us illustrate this point with a numerical example. For the two procedures to have the same power $1 - \beta$ at $\theta_3 = \delta$, the two independent trials with type I error α must each have power $\sqrt{1-\beta}$ at $\theta_3 = \delta$. Suppose that we have normal response with unit variance of the type described in Section 5.2.3, and that one phase III trial with type I error 0.0005 and 90% power at $\theta_3 = 0.2$ will be run. The sample size required for 90% power at $\theta_3 = 0.2$ then equals $n_3 = 1046$ per treatment group. If this sample size is instead equally divided between two independent trials, each with 523 patients per group, these trials would for type I error of $\sqrt{0.0005}$ each have 89% power at $\theta_3 = 0.2$, which gives about 79% probability of both trials achieving statistical significance. It would of course be possible to increase the sample size so that each trial has power $\sqrt{0.9}$ at $\theta_3 = 0.2$, to achieve an overall power of 90%, but such an approach would still be inefficient compared to running one trial with $n_3 = 1046$ patients per treatment group

and type I error probability 0.0005.

5.2.5 Utility function

An important step in decision analysis is to define a utility function. We shall be using a utility function that describes the monetary gain that can be obtained from the sales of a drug, while also taking into account the costs of running the clinical trials needed to get a drug approved. Our model thus takes the perspective of the trial sponsor, for example a pharmaceutical company, and seeks to maximise the expected profits that the drug can deliver. It is however important to note that other views, such as the patient and public health perspectives, are also important. Burman et al. (2007) point out that ethics must be given priority over profit, and several authors have proposed to optimise utilities that are not based on monetary gain. When solving the problem of optimising the phase III sample size for a given prior distribution for the treatment effect, Gittins and Pezeshk (2000b) consider both a public health benefit function, and a gain function based on future sales of the drug. The public health benefit function is assumed to be proportional to the treatment effect, θ_3 in our notation. Anscombe (1963) introduced the so-called horizon problem, where the results of a clinical trial are used to select one of two treatments, that will be used to treat M future patients with the same condition. Here, the objective is to minimise the expected number of patients that are treated with the inferior drug, taking into account both the patients treated in the trial and future patients. Eales and Jennison (1992) studied this problem in the context of a group sequential test. Let $E(N)$ be the expected number of patients who receive the inferior treatment in the group sequential trial. The expected number of patients who receive the inferior treatment, either in the clinical trial or in the future, will then be $E(N) + p_e M$, where p_e is the probability of choosing the wrong treatment for the future patients. The objective is to minimise the expected number of patients who receive the wrong treatment, taking into account both the patients in the clinical trial and future patients.

The utility function in our Bayes problem includes a start-up cost a_2 and a cost per patient c_2 in the phase II trial, a start-up cost a_3 and a cost per patient c_3 in the phase III trial, and a gain g obtained for a statistically significant result, $p < 0.0005$ one-sided, in the phase III trial. We shall be focusing on a constant g in our numerical examples, while acknowledging that a more sophisticated model for g may be useful in some situations. A very natural extension, that will considered in Section 5.6, is to let g depend on the phase III sample size n_3. It is then implicitly assumed that the duration of phase III increases with increasing n_3, which gives a shorter remaining patent life once the drug is approved. Liu et al. (2004) consider this type of model to derive group sequential and adaptive designs for a phase III clinical trial. It would be straightforward to make a similar extension for phase II, and let g depend on both n_2

and n_3. In Section 5.8 we also discuss how the assumption that the gain function g is independent of θ_3 and z_3 can be relaxed.

The process of designing our two trials can be outlined as follows. We first design a two-arm phase II trial comparing a candidate drug to a control, with n_2 patients per treatment group. Once the phase II results are available, a decision about whether to progress to phase III is made. We progress to phase III if $z_2 \geq z_2^\star$ and stop otherwise. If the candidate drug is progressed to phase III, the phase III sample size per group $n_3(z_2, n_2)$ is chosen based the phase II results. After phase III, regulatory approval is achieved if $z_3 \geq z_{1-\alpha}$. Given the threshold $z_{1-\alpha}$ for regulatory approval, defined in Section 5.2.4, and $z_2^\star$, the value of Z_2 that defines the go/no go decision, the utility of the clinical trial program can be written as

$$U(n_2, z_2, n_3, z_3) = -a_2 1_{n_2>0} - c_2 n_2 + 1_{Z_2 \geq z_2^\star} \left\{ g 1_{Z_3 \geq z_{1-\alpha}} - a_3 - c_3 n_3(n_2, z_2) \right\}, \quad (5.12)$$

where n_2 is the within-group sample size of the phase II trial and n_3 is the within-group sample size of the phase III trial. Taking the expectation of equation (5.12), the expected utility that we seek to maximise with respect to n_2, n_3 and $z_2^\star$ can, given our modelling assumptions in (5.3) and (5.11), be written as

$$\begin{aligned} E(U) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\theta_2 d\theta_3 dz_2 dz_3 \pi_{\theta_2,\theta_3}(\theta_2,\theta_3) f_{Z_2|\theta_2}(z_2|\theta_2) \\ &\quad \times f_{Z_3|\theta_3,Z_2}(z_3|\theta_3, z_2)) U(n_2, z_2, n_3(z_2, n_2), z_3). \end{aligned} \quad (5.13)$$

The unconditional probability of rejecting the null hypothesis $\theta_3 \leq 0$, averaged over the prior distribution of the treatment effect before phase III, is sometimes referred to as the assurance. As the pdf of Z_3 depends on θ_3, an integral over the prior distribution of θ_3 before the phase III trial would generally be required to calculate the assurance. Suppose that conditional on $Z_2 = z_2$ and n_2, the posterior for θ_3 after phase II, which is also the prior for θ_3 before phase III, is normally distributed according to

$$\theta_3 \sim N(\mu_2(z_2, n_2), \tau_2^2(n_2)).$$

O'Hagan et al. (2005) show that the assurance can be written as

$$\begin{aligned} \gamma(z_2, n_2, n_3) &= \int_{z_{1-\alpha}}^{\infty} dz_3 \int_{-\infty}^{\infty} d\theta_3 \pi_{\theta_3|Z_2}(\theta_3|z_2) f_{Z_3|\theta_3,Z_2}(z_3|\theta_3, z_2) \\ &= \Phi\left(\frac{\mu_2(z_2, n_2)\sqrt{n_3(z_2, n_2)/(2\sigma_3^2)} - z_{1-\alpha}}{\sqrt{1 + n_3(z_2, n_2)\tau_2^2(n_2)/(2\sigma_3^2)}} \right). \end{aligned} \quad (5.14)$$

In Section 5.8.1 we show how this expression for the assurance γ can be used to simplify (5.13) to obtain

$$\begin{aligned} E(U) &= -a_2 1_{n_2>0} - c_2 n_2 + \int_{z_2^\star}^{\infty} dz_2 f_{Z_2}(z_2) \\ &\quad \times \{\gamma(z_2, n_2, n_3(z_2, n_2))g - a_3 - c_3 n_3(z_2, n_2)\}, \end{aligned} \tag{5.15}$$

where $f_{Z_2}(z_2) = \int_{-\infty}^{\infty} d\theta_2 \pi_{\theta_2}(\theta_2) f_{Z_2|\theta_2}(z_2|\theta_2)$ is the marginal pdf of Z_2.

In practice, it can be difficult to know which value of g to use in (5.15). One interpretation is that g is the expectation of a random variable G, which depends on many uncertain factors, such as unexpected safety problems of the drug that may prevent regulatory approval, or the number of competitor drugs that will be on the market if the drug is approved. Suppose that the trial sponsor's prior beliefs about these uncertain factors can be expressed as prior distributions. The expected value of G can then be obtained by integrating over these prior distributions. Let us illustrate how this could be done with a very simple example. Assume that the prior belief is that independently of the treatment effect, a safety problem that prevents regulatory approval will occur with probability 0.1. Suppose that $g = M$ would be used in the expression for the expected utility, if the prior probability of a safety problem were zero. If the prior probability of a safety problem is 0.1 instead of zero, it is appropriate to replace $g = M$ with $g = 0.9M$.

We consider the design of one phase II trial and one phase III trial. The objective is to design these two trials in a way that maximises the expected utility defined in (5.15). We thus seek to find the optimal phase II sample size, $n_2^\star$, as well as the threshold, $z_2^\star$, for the go/no go decision rule for progressing to phase III based on phase II data. It follows from equations (5.4) and (5.8) that μ_2, the posterior mean of θ_3 after phase II, is linear in z_2. Hence, given prior information and phase II sample size, there is a one-to-one correspondence between z_2 and μ_2. Consequently, we shall sometimes be referring to the cut-off as $\mu_2^\star$, rather than $z_2^\star$.

In some cases the model will be evaluated in the context of a fixed phase III sample size n_{3f}. We shall however also consider finding $n_2^\star$ and $\mu_2^\star$ when n_3 is not taken to be fixed, but is a function $n_3(z_2, n_2)$ of the phase II trial results. Derivation of how to optimally choose n_3 is then also part of the optimisation procedure. In addition, we will discuss how the properties of the model change if phase III is made group sequential.

5.3 Optimisation of phase II given fixed sample phase III design

5.3.1 Basic assumptions

In this section we consider the situation when the phase III sample size n_3 is fixed and the phase II sample size n_2 is optimised given this phase III design. Although not a crucial assumption for the model to work, we consider trials with two treatment groups of equal size, in both phase II and phase III. All the sample sizes referred to are per-group sample sizes, so the total sample size will be $2n_2$ in the phase II trial and $2n_3$ in the phase III trial.

Two different approaches to choosing a fixed phase III sample size n_3 will be considered in this section. In neither case is n_3 allowed to depend on the results from phase II. The first approach will be to assume that the phase III sample size takes a fixed value n_{3f}. This value may be motivated by achieving a certain power at a given effect size that is thought be of clinical relevance. There may also be other reasons for n_3 taking a fixed value, for example the requirement to collect a minimal amount of safety data. The second approach is to choose the phase III sample size $n^\star_{3f}$ that maximises the expected utility. This phase III sample size is optimal within the class of designs where the phase III sample size is not allowed to depend on the phase II results.

It is clear that the phase II results will be available before starting the phase III trial. So one may ask if there is a practical reason to insist on searching for $n^\star_{3f}$, without using the knowledge gained from the phase II data. One answer is that for planning purposes, it can be useful to have a phase III sample size in mind before the phase II results are available. With a working assumption about what the phase III sample size will be, planning for how to deal with issues such as patient recruitment and drug supply can start. There is also a theoretical interest in assessing the benefit of adaptively choosing the phase III sample size based on phase II data. This assessment bears similarities with the comparison of adaptive and non-adaptive group sequential tests for superiority and non-inferiority, addressed in Chapter 2. We shall come back to this question in Section 5.5, where the phase III sample size is allowed to depend on phase II data.

Even though the focus will be on general conclusions, it is useful to assume some numerical values for the different parameters included in our model. We can without loss of generality set the within-group sample variances in both phase II and phase III to unity, so $\sigma_2^2 = \sigma_3^2 = 1$. For simplicity, we assume no start-up costs for now, i.e. $a_2 = a_3 = 0$. The three parameters c_2, c_3 and g, which all have the same unit, can be described by two degrees of freedom. We can write the expected utility as $c_3 \times f(c_2/c_3, g/c_3)$, so to maximise the expected utility it is enough to specify the

ratios c_2/c_3 and g/c_3. Hence, we can without loss of generality set $c_3 = 1$, and the expected utility will be displayed for this case in our numerical examples. The optimal choices of n_2, n_3 and $z_2^\star$ will however apply also for other choices of c_3, provided that c_2/c_3 and g/c_3 remain fixed.

It should be possible to determine c_2/c_3 with reasonable accuracy, while g/c_3 may be more difficult to estimate precisely. We shall assume a constant ratio $g/c_3 = 12000$, so the profit from a phase III trial is 12000 times the cost of one phase III subject per treatment arm. Our choice of $g/c_3 = 12000$ is mainly for illustrative purposes. It would appear to be a reasonable assumption and lies within the range published by Gittins and Pezeshk (2000a), where Bayesian decision analysis is used to design phase III trials of six different drug projects. In a practical situation it is important to consider the uncertainty around each parameter and perform sensitivity analyses, to assess how robust the model is to misspecification of parameters, such as for example g/c_3.

5.3.2 Proportion of resources in phase II and phase III

Let us now consider the problem of maximising the expected utility in (5.15), for the case when n_3 is not allowed to depend on the phase II results. We seek to find $n_2^\star$, both for the case when $n_3 = n_{3f}$ and for $n_3 = n_{3f}^\star$. Let us first consider the case when n_{3f} is not optimised, but is instead based on other considerations, for example power requirements. Suppose that for the condition that is being investigated, it is found appropriate to set power at a clinically relevant difference of $\theta_3 = \delta$. To obtain 90% power at $\delta = 0.2$, we set

$$n_{3f} = \frac{2\sigma_3^2(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta))^2}{\delta^2} = 1046.$$

We further assume

$$(\theta_2, \theta_3) \sim N\left((0,0), \begin{pmatrix} 0.04 & 0.04r \\ 0.04r & 0.04 \end{pmatrix}\right), \tag{5.16}$$

so that power is set one standard deviation, of the prior distribution of θ_3, away from the prior mean of θ_3. The assumption of $\mu_1 = 0$ may be regarded as a bit pessimistic, in particular for comparisons against placebo. It is however an interesting case to study, as before phase II it is far from certain whether the drug will progress to phase III, and the phase II trial can play an important role in the go/no go decision. A more optimistic prior distribution for θ_3 will be considered in Section 5.6. A prior distribution for θ_3 centred at $\mu_1 = 0$ may be particularly realistic in a development program where the experimental drug is compared to an active control. If superiority over the active control is required for regulatory approval, this fits very nicely within our current framework. Our model could also handle the situation when establishing non-inferiority for a pre-

specified non-inferiority margin of δ_N is deemed sufficient. In the latter case, the null hypothesis for θ_3 in phase III would be $\theta_3 \leq -\delta_N$, rather than $\theta_3 \leq 0$.

When evaluating the amount of resources spent in phase II and phase III respectively, it is natural to consider the sample size ratio n_2/n_3. Apart from the proportion of patients, it is also of interest to assess the relative costs of the two phases. Denote the investment in phase II by

$$C_2 = a_2 + c_2 n_2$$

and the investment in phase III by

$$C_3 = a_3 + c_3 n_3.$$

We define the phase II/phase III investment ratio as

$$\frac{C_2}{C_3} = \frac{a_2 + n_2 c_2}{a_3 + n_3 c_3}. \tag{5.17}$$

Once we know the choices of phase II and phase III sample sizes, it is straightforward to calculate the investment and sample size ratio for different choices of c_2/c_3.

We have used the methods described in Section 5.8 to solve this decision problem, for both $n_3 = 1046$ and $n_3 = n_{3f}^{\star}$. The problem has been solved for various values of the correlation r and three specific values of the ratio c_2/c_3, assuming $g/c_3 = 12000$ and the prior distribution in equation (5.16). The optimal phase II and phase III sample sizes, sample size ratios and investment ratios are shown in Figure 5-2. The results are qualitatively similar for the two ways of choosing phase III sample size that have been considered. We see in Figure 5-2 that $n_2^{\star}$ increases monotonically with increasing r. This is not surprising, as more information about θ_3 is obtained for larger r. It is perhaps less obvious that there would be a cut-off $r^{\star}$, so that $n_2^{\star} = 0$ for $r < r^{\star}$. The implication of the latter is that we would move directly to phase III, without first running a phase II trial. The reason for this discontinuity in the choice of n_2 will be further discussed in Section 5.4, where we focus on the value of information and the decisions that are based on the information gathered in phase II.

In practice it can be difficult to know which value to use for the correlation r in the prior distribution for θ_2 and θ_3. Figure 5-2 can however be used to give an assessment of whether a certain pair of planned sample sizes n_2 and n_3 seem reasonable. Suppose that c_2/c_3 can be estimated with reasonable accuracy. If $c_2/c_3 \geq 0.2$, we can for example see that a plan with $n_2/n_3 \geq 0.4$ does not make sense, unless the correlation between θ_2 and θ_3 is close to one.

The case $r = 1$ and $c_2 = c_3 = 1$, which can be observed in Figure 5-2, is of special interest as it can be interpreted as running a pre-study with the same endpoint that is

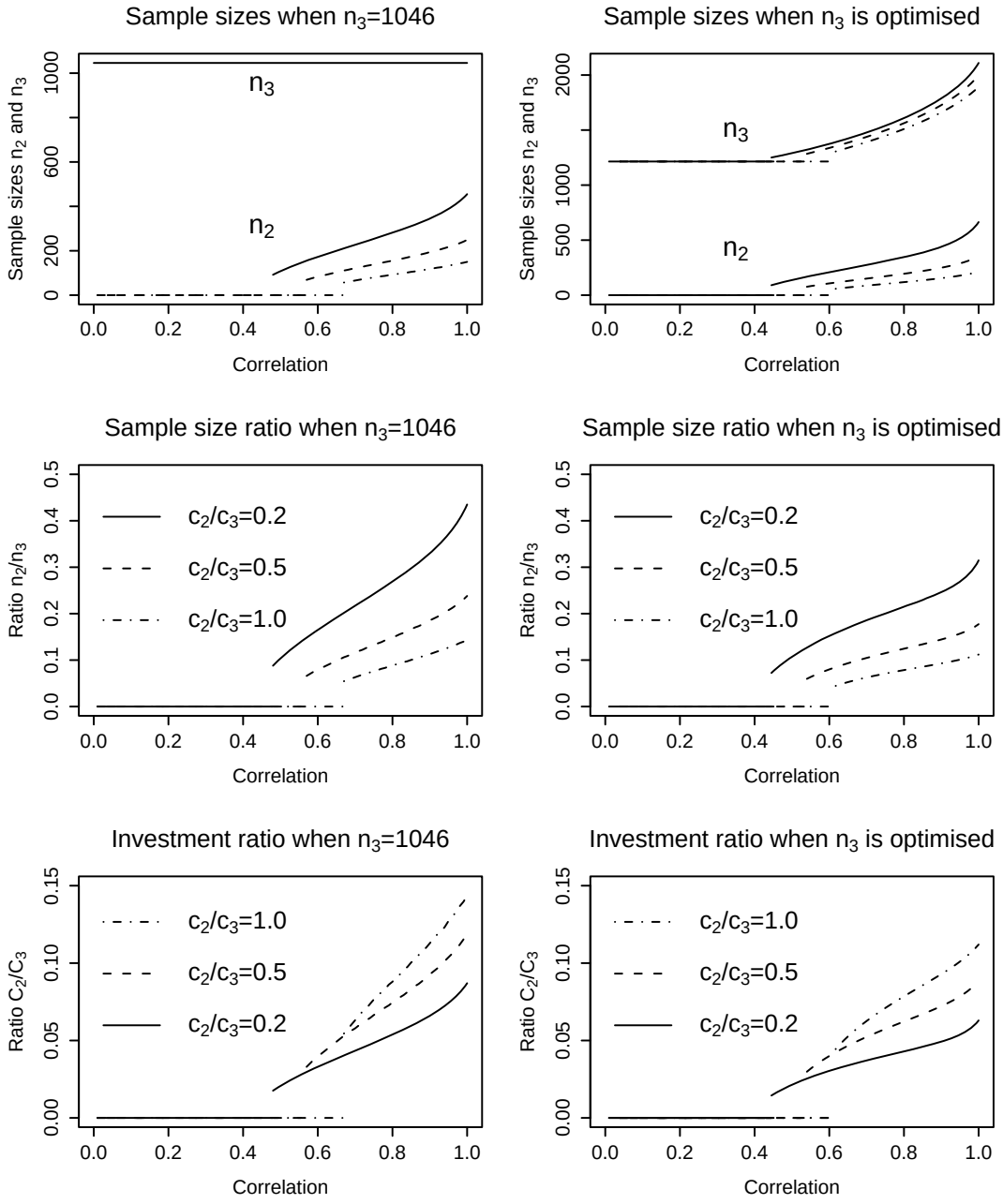

Figure 5-2: Optimal phase II and phase III sample sizes, sample size ratios and investment ratios for different values of the correlation r and different cost ratios c_2/c_3. In the three panels to the left the phase III sample size is fixed at $n_3 = 1046$, while in the three panels to the right the phase III sample size is chosen to maximise the expected utility. The designs have been optimised for $g/c_3 = 12000$ and prior distribution for (θ_2, θ_3) according to (5.16).

later used in phase III. The phase II observations are cheap and can provide information about θ_3, so one may ask why the phase II sample size is not higher for $r = 1$. The reason is that the phase II observations are not used in the final hypothesis test in phase III.

So why are some resources spent in phase II rather than phase III, when the observations in phase II are not included in the final analysis? Another option would be to instead use these resources in phase III. The latter would appear to be advantageous, as the observations then can also contribute to increasing the power to reject the null hypothesis in the confirmatory trial. The benefit of the observations in phase II is that after phase II, you get a chance to stop development of drugs that do not appear to be efficacious. Since phase III is a fixed sample trial rather than group sequential, this is not possible in phase III.

A final observation is related to the seamless phase II/III trials that were briefly discussed in Section 5.1. We said that we would not consider seamless phase II/III trials in this chapter, but note that such designs would be particularly advantageous if the same endpoints could be used in phase II and phase III.

5.3.3 Sensitivity to specification of gain function

We have already mentioned that the gain g may be difficult to establish with high accuracy and that in practical situations, it is important to assess how robust the proposed design is to uncertainties about g and the ratio g/c_3. In Figure 5-3 we set $c_2/c_3 = 0.2$ and show how the investment in phase II and the resource allocation between phase II and phase III depend on g/c_3.

Let us first consider the case when n_3 is fixed, based on power or other considerations, and does not change depending on g/c_3. For small g/c_3, the case for investing in phase II increases for increasing g/c_3. But there comes a point when g/c_3 is so large that the phase II trial is unlikely to change the decision about whether to run the phase II trial. The phase II trial then becomes less important for the go/no go decision, as clearly illustrated in the curve for $r = 0.7$, where $n_2^\star = 0$ for large enough g/c_3. We would expect to eventually see the same behaviour for $r = 0.8$ and $r = 0.9$, even though these cases are not shown in Figure 5-3. It might be expected that a small investment in phase II should help, as it saves the cost of phase III trials that are unlikely to give positive outcomes. For low enough $Z_2 = z_2$, the drug would not be progressed to phase III. But as g/c_3 increases, the threshold $z_2^\star$ for progress to phase III shifts in the negative direction, and the probability that phase II makes a difference gets smaller.

The situation is different for $n_3 = n_{3f}^\star$ when, as illustrated in the right panel of Figure 5-3, $n_{3f}^\star$ varies with g/c_3. Both $n_2^\star$ and $n_{3f}^\star$ are monotonically increasing in g/c_3. For small g/c_3, the phase II sample size initially increases more rapidly than n_3. As g/c_3

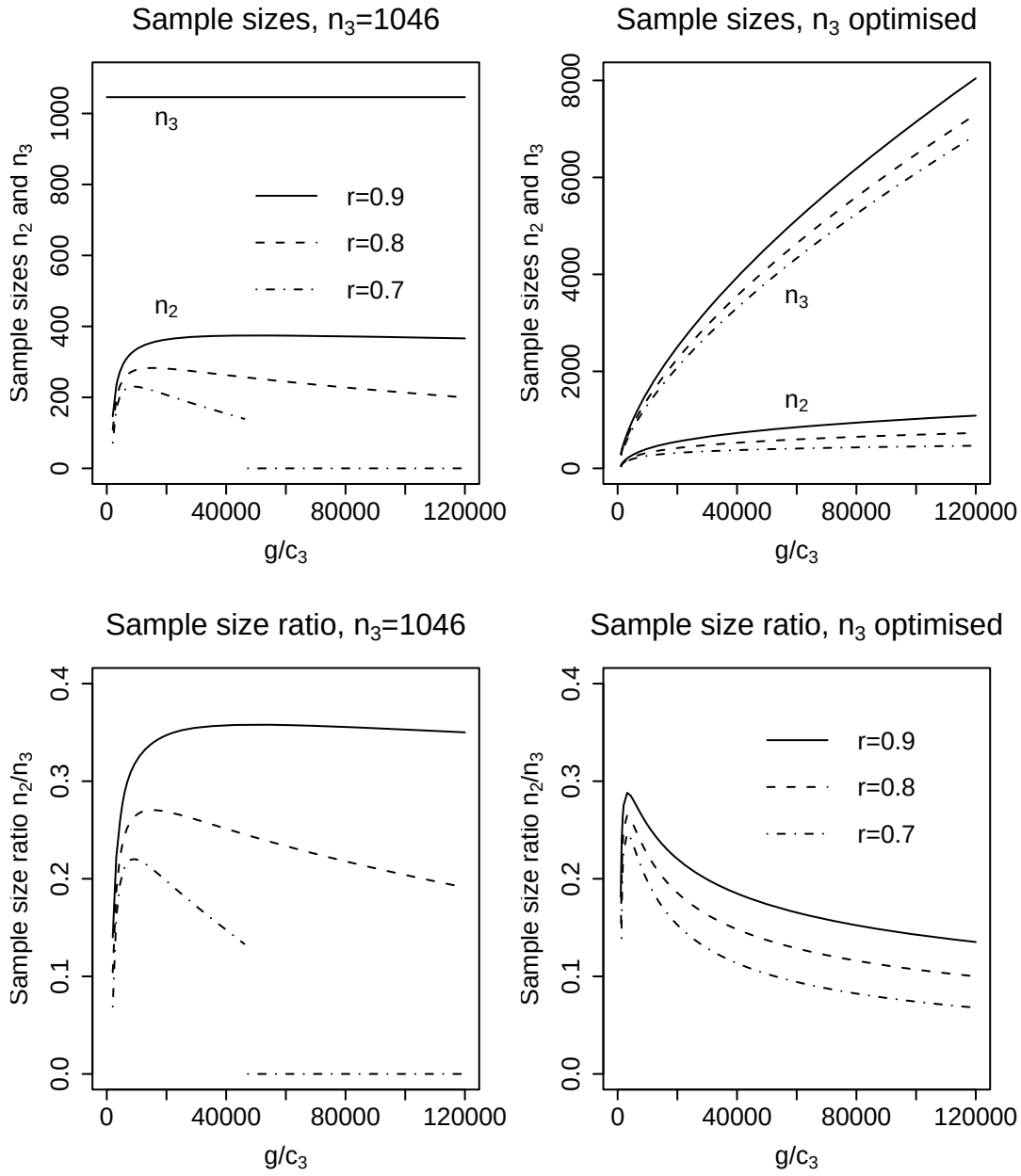

Figure 5-3: Optimal phase II and phase III sample sizes and sample size ratios, for different values of the correlation r and ratios g/c_3. In the three panels to the left the phase III sample size is fixed at $n_3 = 1046$, while in the three panels to the right the phase III sample size is chosen to maximise the expected utility. The designs have been optimised for $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16).

increases however, it becomes more important to allocate resources to the phase III trial and ensure success there. The ratio $n_2/n_{3f}^\star$ thus increases initially, to a level of about 0.25-0.3, but soon starts to increase. Unlike when $n_3 = 1046$, the optimal phase II sample size does not go down to zero. We believe that this is because the investment in phase III continues to increase, with increasing g/c_3. Even though the probability of not running phase III decreases, the associated cost of phase III is getting very high. The go/no go decision then becomes more important, which motivates an increasing investment in phase II.

5.3.4 Choice of phase II sample size and impact on expected utility

We have studied how the optimal phase II sample size $n_2^\star$ depends on various combinations of c_2/c_3, g/c_3 and r, as well as different ways of choosing the phase III sample size. Let us now consider an example of how the expected utility and probability of success (PoS) actually depend on the choice of phase II sample size. We choose to focus on $r = 0.8$ and $r = 0.45$, as they represent two qualitatively different cases. In the former, the biomarker can provide useful information about the phase III endpoint while in the latter, sampling in phase II is a rather inefficient way of learning about θ_3. We further assume $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution according to equation (5.16). We shall refer to the case with

$$\begin{aligned} g/c_3 &= 12000 \\ c_2/c_3 &= 0.2 \\ (\theta_2, \theta_3) &\sim N\left((0,0), \begin{pmatrix} 0.04 & 0.04r \\ 0.04r & 0.04 \end{pmatrix}\right) \end{aligned} \tag{5.18}$$

as our core example, and return to it at various points during this chapter.

Let us first consider our core example for the case when the correlation $r = 0.8$. The upper left panel of Figure 5-4 shows the expected utility for different values of n_2. When $r = 0.8$, running a phase II trial with sample size $n_2 = n_2^\star$ gives substantial benefits, compared to when $n_2 = 0$. For the two different ways of choosing n_{3f} that we have considered, there is a value $n_2^\star$ that maximises the expected utility. It is noteworthy that the expected utility actually decreases slightly in n_2 for very small n_2, and reaches a local minimum before it starts to increase. There is then a second change in the sign of the first derivative at the local maximum of our expected utility, which occurs at the optimal phase II sample size $n_2^\star$. Compared to the option of not running a phase II trial, shown by $n_2 = 0$, there is a substantial benefit in running a phase II trial with $n_2 = n_2^\star$.

The situation is different when we consider another variation of our core example, with $r = 0.45$ instead of $r = 0.8$. For $n_{3f} = 1046$ it is actually beneficial to move

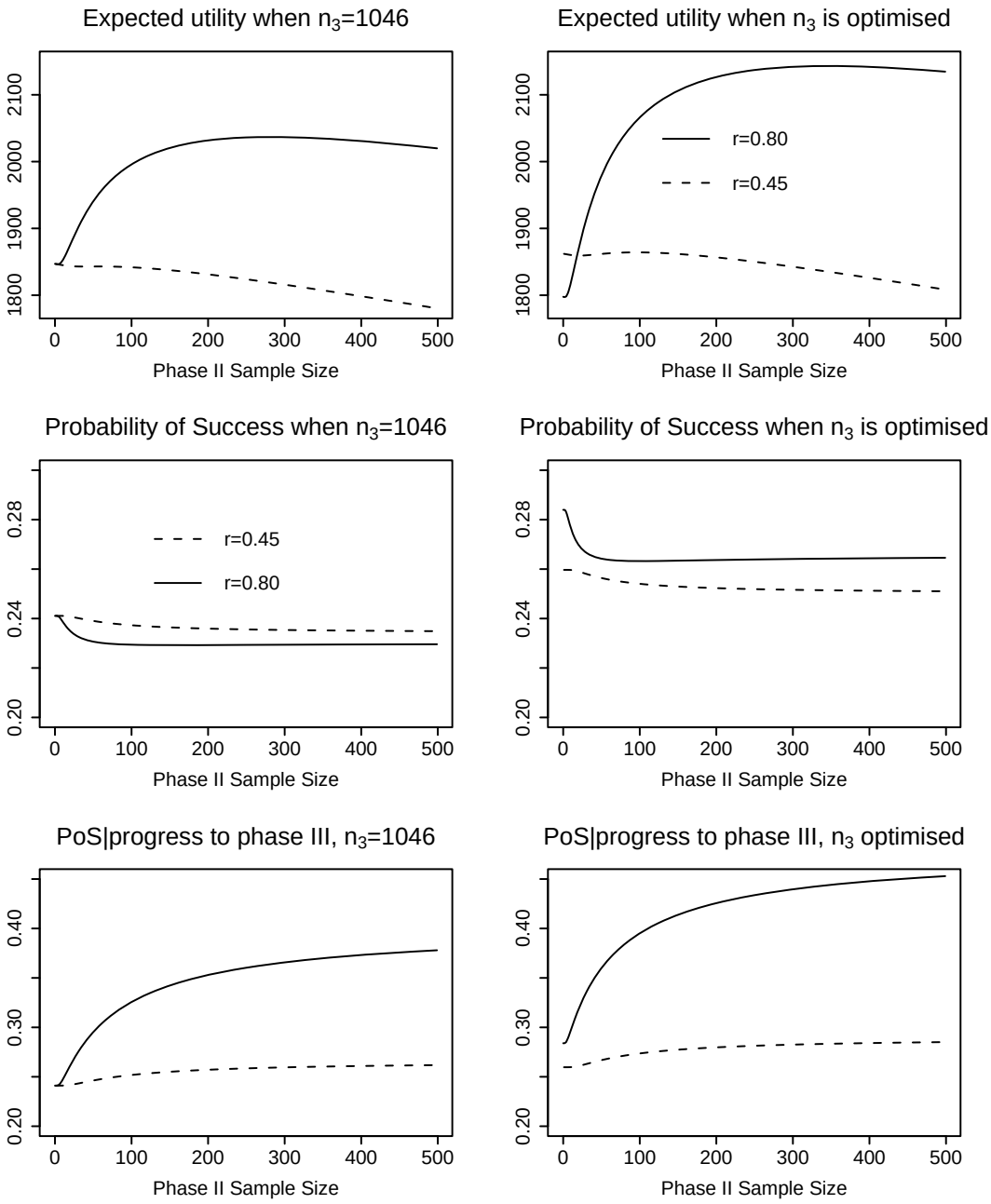

Figure 5-4: Expected utility and probability of success (PoS), conditional and unconditional on progress to phase III, for $r = 0.8$ (solid line), $r = 0.45$ (dashed line) and different values of n_2. The panels to the left show results for $n_3 = 1046$ and the panels to the right show results for $n_3 = n_{3f}^{\star}$. The designs have been optimised for our core example, i.e. $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16).

directly to phase III, i.e. setting $n_2 = 0$. There is a local maximum for just under 60 observations per treatment group, but the expected utility at this local maximum is smaller than for $n_2 = 0$. For $n_{3f}^{\star} = 1254$, it is optimal to run a phase II trial with just under 100 patients per treatment group. The impact of the phase II trial on the expected utility is however very modest, as illustrated in the dashed line in Figure 5-4. In fact, we believe the most important conclusion for $r = 0.45$ to be that it is not very helpful to run a phase II trial when the correlation is this low.

Figure 5-4 also shows the probability of success, conditionally and unconditionally on progressing to phase III. For the cases that we have considered, there is typically an initial small decrease in the unconditional PoS, as some paths to a successful result in phase III are now stopped after the phase II trial. While the impact on the unconditional PoS is minor, there is a considerable increase in the PoS conditional on progressing to phase III. This increase is however much smaller for $r = 0.45$ than for $r = 0.8$. Given that the phase III sample size is not allowed to depend on the phase II results, the role of the phase II trial is to identify whether a treatment should be brought forward to phase III. As the probability of success in phase III is substantially increased, it would appear that for $r = 0.8$, the phase II trial is successful in identifying treatments that should not be brought forward. This should be reassuring to investigators and trial sponsors, who are keen to avoid costly failures at the last hurdle.

5.3.5 Threshold for progress to phase III

When phase II is not used to choose the phase III sample or for dose-finding, its main purpose is to enable a well informed go/no go decision. It is thus of interest to assess how $\mu_2^{\star}$, the minimum of the posterior mean of θ_3 after phase II that is required for progress to phase III, depends on the other parameters in the model. Our decision rule is to progress the drug to phase III if $\mu_2 \geq \mu_2^{\star}$ and stop development otherwise. It should be recognised that in practice, many other aspects will have an impact on the go/no go decision. The size of the phase III investment, the degree to which other treatments for the condition exist and the safety profile of the compound are examples of other features that are likely to be important, resulting in a multi-dimensional problem. Ideally, these additional features should be included in the model. Nevertheless, our simplified model provides a useful quantitative assessment of how to proceed.

For a fixed phase III sample size n_3, we can use the assurance γ, as defined by O'Hagan et al. (2005), to calculate $\mu_2^{\star}$. For the expected utility, conditional on $Z_2 = z_2$ and n_2, to be positive it is necessary that

$$\gamma g \geq a_3 + c_3 n_3.$$

The limiting case, when the expected reward of running a phase III trial equals the cost, occurs when

$$\gamma = \frac{a_3 + c_3 n_3}{g}. \tag{5.19}$$

O'Hagan et al. (2005) show that if

$$\theta_3 | Z_2 = z_2 \quad \sim \quad N(\mu_2, \tau_2^2),$$

the assurance γ can be calculated directly from the standard normal cumulative distribution function according to

$$\gamma = \Phi\left(\frac{\mu_2 \sqrt{n_3/(2\sigma_3^2)} - z_{1-\alpha}}{\sqrt{1 + n_3 \tau_2^2/(2\sigma_3^2)}} \right).$$

The next step is to insert this expression for γ into (5.19) and solve for μ_2. We then find that the threshold $\mu_2^\star$ must satisfy

$$\mu_2^\star = \frac{z_{1-\alpha} - \Phi^{-1}(1 - (a_3 + c_3 n_3)/g)\sqrt{1 + n_3 \tau_2^2/(2\sigma_3^2)}}{\sqrt{n_3/(2\sigma_3^2)}} \tag{5.20}$$

We would intuitively expect to progress only if μ_2 is positive, but will now show that this property holds only if certain conditions apply. It follows from (5.20) that $\mu_2^\star = 0$ if and only if

$$\begin{aligned} \Phi^{-1}(1 - (a_3 + c_3 n_3)/g) &= \frac{z_{1-\alpha}}{\sqrt{1 + n_3 \tau_2^2/(2\sigma_3^2)}} \\ \Leftrightarrow \frac{g}{a_3 + c_3 n_3} &= \frac{1}{\Phi(-z_{1-\alpha}/\sqrt{1 + n_3 \tau_2^2/(2\sigma_3^2)})} \end{aligned} \tag{5.21}$$

If $\tau_2^2 = 0$, (5.21) holds when

$$\frac{g}{a_3 + c_3 n_3} = \frac{1}{\alpha}.$$

In our model, with a one-sided type I error of 0.0005, a ratio $g/(a_3 + c_3 n_3) \geq 2000$ would thus be necessary to obtain a negative threshold $\mu_2^\star$ in this case. Hence, $n_3 \leq 6$ would be required in our core example with $g/c_3 = 12000$ and $a_3 = 0$. This is re-assuring, as a trial with so few observations is unlikely to be convincing to regulatory authorities and the wider scientific community.

Let us now consider the case $\tau_2^2 > 0$. For $\tau_2^2 > 0$,

$$\frac{g}{a_3 + c_3 n_3} \geq \frac{1}{\alpha}$$

is a sufficient but not a necessary condition for obtaining $\mu_2^\star < 0$. In most practical situations we will have $n_3\tau_2^2 >> 2\sigma_3^2$, so for $\mu_2^\star$ to be positive we must have

$$\frac{g}{a_3 + c_3 n_3} < \frac{1}{\alpha}.$$

The situation is different from when $\tau_2^2 = 0$, as the probability mass for positive values of θ_3 may make it worthwhile to run the phase III trial.

In summary, we have shown that $\mu_2^\star$ can be found analytically when n_3 is not allowed to depend on the phase II sample size. The threshold depends on the gain function g, the cost of the phase III trial, the prior distribution for θ_3 before phase III and the threshold $z_{1-\alpha}$ for regulatory approval. If the assumption that n_3 is fixed is relaxed, the threshold becomes smaller than if n_3 is fixed, as there are more options for how to design phase III. The situation when n_3 is allowed to depend on the phase II results is further discussed in Section 5.5.2.

5.4 The value of information

5.4.1 Information about θ_3

Before moving on to the more complicated problem of optimising the phase III design based on phase II data, it may be useful to take a step back and think about the role of the phase II trial in our model. To this end, we will be following the framework described in Section 5.2.3. According to our model a certain investment is made in phase II, with the hope to gain information about the clinical endpoint used in phase III. Assuming $a_2 = 0$, we invest $C_2 = c_2 n_2$ in the phase II trial, where we wish to learn as much as possible about θ_3, the mean of the phase III endpoint. Learning in phase II can be particularly beneficial if the information is cheaper than purchasing the same information in phase III. The observations in phase II have the possibility to contribute to the go/no go decision before phase III, while the observations in phase III are used for the final hypothesis test in phase III.

Before phase II, θ_3 has prior variance τ_1^2, while after phase II the posterior variance of θ_3 is denoted τ_2^2. It is defined in equation (5.9) and can be written as

$$\tau_2^2 = \tau_1^2(1 - r^2(1 - t_2^2/t_1^2)).$$

Taking the posterior information as the inverse of the posterior variance of θ_3, it is natural to focus on the information gain

$$\frac{1}{\tau_2^2} - \frac{1}{\tau_1^2} = \frac{1}{\tau_1^2(1 - r^2(1 - t_2^2/t_1^2))} - \frac{1}{\tau_1^2}. \tag{5.22}$$

For a given phase II sample size n_2, we can insert the expression for t_2^2 in (5.5) into equation (5.22) to obtain

$$\begin{aligned}\frac{1}{\tau_2^2} - \frac{1}{\tau_1^2} &= \frac{r^2(1 - t_2^2/t_1^2)}{\tau_1^2(1 - r^2(1 - t_2^2/t_1^2))} \\ &= \frac{r^2 n_2 t_1^2/(2\sigma_2^2)}{\tau_1^2(1 + (1 - r^2)n_2 t_1^2/(2\sigma_2^2))} \\ &= \frac{r^2 C_2 t_1^2/(2\sigma_2^2 c_2)}{\tau_1^2(1 + (1 - r^2)C_2 t_1^2/(2\sigma_2^2 c_2))},\end{aligned} \tag{5.23}$$

as a measure of the information about θ_3 obtained due to the phase II trial. We see in (5.23) that if it were possible to sample a biomarker with $r = 1$, it would simplify to $C_2 t_1^2/(2\sigma_2^2\tau_1^2 c_2)$. The information about θ_3 obtained from such a biomarker increases linearly in C_2, as the term in the denominator that involves C_2 disappears. Another benefit with using the phase III endpoint in phase II would be that combining data from the two phases could be more easily accomplished. In many situations it is however rather unrealistic to have a biomarker whose mean has correlation $r = 1$ with θ_3. Even if it is possible to collect the phase III endpoint in phase II, there are likely to be other issues, such as differences in patient population, that give a correlation below one. A more realistic option may be to measure the same variable in phase II as in phase III, but with a much shorter follow-up time. The correlation will then be below one, but the need for a high correlation can be balanced with the time and cost of collecting the data.

Equation (5.23) can be written as

$$\frac{1}{C_2}\left(\frac{1}{\tau_2^2} - \frac{1}{\tau_1^2}\right) = \frac{r^2 t_1^2/(2\sigma_2^2 c_2)}{\tau_1^2(1 + (1 - r^2)C_2 t_1^2/(2\sigma_2^2 c_2))}. \tag{5.24}$$

We can interpret the left-hand side of equation (5.24) as the information obtained per unit cost. Initially, when $C_2 \approx 0$, the right-hand side of equation (5.24) is approximately

$$\frac{r^2 t_1^2/(2\sigma_2^2 c_2)}{\tau_1^2}$$

If an investment of the same size is made in phase III, the information obtained per unit cost can be written as

$$\frac{1}{2\sigma_3^2 c_3}.$$

We can compare these two expressions to get an idea about the information gained per unit cost, depending on whether it is purchased in phase II or phase III. The two expressions are however not directly comparable, as the information obtained have

different roles in our model. The information obtained in phase II can be used to make a decision about progression to phase III, but is not included in the final hypothesis test in phase III. The role of the phase III observations is the opposite. They are included in the final hypothesis test but, since phase III is a fixed sample trial, cannot be used to make a decision to stop development of the drug.

5.4.2 Choice of biomarker

We continue to follow the framework described in Section 5.2.3 and further discussed in Section 5.4.1. Suppose that we are considering two biomarkers, with means θ_{21} and θ_{22} respectively, for potential use in a phase II clinical trial. In some situations, it may be possible to use both biomarkers, in which case it would be necessary to model the correlation structure between θ_3, θ_{21} and θ_{22}. We shall however consider the situation when one of the two biomarkers must be chosen, so for each biomarker we need to specify the parameters introduced in the model description in Section 5.2.3. Let θ_{21} and θ_{22} have correlations r_1 and r_2 with θ_3 respectively, where $0 < r_1 < r_2$. The costs per observation are c_{21} and c_{22}, with $c_{21} < c_{22}$. It is then far from obvious which biomarker will be more efficient for learning about θ_3, as θ_{22} is more strongly correlated with θ_3, while the cost per observation is lower for biomarker 1. Suppose θ_{21} and θ_{22} are the means of the same variable, but measured after a different follow-up time. It is then reasonable to expect that if θ_{22} represents the mean after a longer follow-up than θ_{21}, it will be more costly to measure, but also more strongly correlated with θ_3. So this is an example of a situation where it is not straightforward to choose which biomarker to use.

When analysing which biomarker to use in the phase II trial, it is of interest to find the levels of phase II investment, where one of the biomarkers is to be preferred over the other. Assume that θ_{21} and θ_{22} have prior variances t_{11}^2 and t_{12}^2, while the response variances are σ_{21}^2 and σ_{22}^2. It then follows from (5.23) that the two biomarkers give equal amount of information if

$$\begin{aligned} & \frac{r_1^2 C_2 t_{11}^2/(2c_{21}\sigma_{21}^2)}{\tau_1^2(1+(1-r_1^2)C_2 t_{11}^2/(2c_{21}\sigma_{21}^2))} \\ = \; & \frac{r_2^2 C_2 t_{21}^2/(2c_{22}\sigma_{22}^2)}{\tau_1^2(1+(1-r_2^2)C_2 t_{21}^2/(2c_{22}\sigma_{22}^2))} \\ \Leftrightarrow \; & C_2 = 0 \text{ or } C_2 = C_2^\star = \frac{2(c_{22}\sigma_{22}^2 r_1^2 t_{12}^2 - c_{21}\sigma_{21}^2 r_2^2 t_{11}^2)}{(r_2^2 - r_1^2)t_{11}^2 t_{12}^2}. \end{aligned} \tag{5.25}$$

Solutions where $C_2 < 0$ are of no interest, but the knowledge that learning about θ_{21} delivers more information if

$$0 < C_2 < C_2^\star$$

provides some useful insights. For phase II investments smaller than $C_2^\star$, θ_{21} will dominate, while for larger investments than $C_2^\star$, θ_{22} is to be preferred.

It is natural to ask how much cheaper biomarker 1 must be, to compensate for the fact that the mean of biomarker 2 is more strongly correlated with θ_3. We can solve equation (5.25) for c_{21} to find that running a phase II trial with biomarker 1 as opposed to biomarker 2, for a fixed level of investment C_2, gives more information about θ_3 if

$$c_{21} < \frac{2c_{22}r_1^2\sigma_2^2t_{11}^2 + C_2t_{12}^2t_{11}^2(r_1^2 - r_2^2)}{2r_2^2\sigma_1^2t_{12}^2}. \tag{5.26}$$

The right hand side of the inequality in (5.26) thus depends linearly on C_2, the level of phase II investment. Since $r_1 < r_2$, the most favourable situation for biomarker 1 occurs when $C_2 = 0$. In the limit, as $C_2 \to 0^+$, we find that the increase in information is higher for θ_{21} only if

$$\frac{c_{21}}{c_{22}} < \frac{r_1^2t_{11}^2\sigma_{22}^2}{r_2^2t_{12}^2\sigma_{21}^2}. \tag{5.27}$$

Unless (5.27) holds, we can directly dismiss a biomarker with a mean that is less strongly correlated with θ_3, as the purchase of information will only become less favourable as C_2 increases. As $C_2 \to \infty$, we find that biomarker 1 is more efficient if

$$\frac{r_1^2}{1 - r_1^2} > \frac{r_2^2}{1 - r_2^2},$$

which for $r_1 \neq 1$ and $r_2 \neq 1$ is equivalent to

$$r_1^2 > r_2^2.$$

If one of the biomarkers has correlation $r = 1$ and the other not, the biomarker with correlation $r = 1$ will be more efficient as $C_2 \to \infty$. It remains to consider the case $r_1 = r_2 = 1$, in which case biomarker 1 is more efficient as $C_2 \to \infty$, if

$$\frac{c_{21}\sigma_{21}^2}{t_{11}^2} < \frac{c_{22}\sigma_{22}^2}{t_{12}^2}.$$

Figure 5-5 shows the information obtained for θ_3 for three different biomarkers, with means θ_{21}, θ_{22} and θ_{23}. In some situations it might be possible to use more than one biomarker in the phase II trial, but that is not our focus here. The condition in (5.25) can be used to decide which biomarker to use in a certain situation. The three biomarkers are assumed to have different costs per observation, but the same sample variance σ_2^2. The means θ_{21}, θ_{22} and θ_{23} have the same prior variance t_1^2, but different correlations with θ_3. To find the roots of C_2 where two biomarkers give equal amount

of information for θ_3 we can use (5.25), which simplifies to

$$C_2^\star = \frac{2\sigma_2^2(c_{22}r_1^2 - c_{21}r_2^2)}{t_1^2(r_2^2 - r_1^2)}.$$

We can for $t_1^2 = 0.04$ and $\sigma_2^2 = 1$ predict that the curves in Figure 5-5 will cross at investment levels of

$$C_2 = \frac{2 \times 1^2(0.5 \times 0.7^2 - 0.2 \times 0.8^2)}{0.04(0.8^2 - 0.7^2)} = 39.00 \tag{5.28}$$

$$C_2 = \frac{2 \times 1^2(0.7^2 - 0.2 \times 0.9^2)}{0.04(0.9^2 - 0.7^2)} = 51.25 \tag{5.29}$$

$$C_2 = \frac{2 \times 1^2(0.8^2 - 0.5 \times 0.9^2)}{0.04(0.9^2 - 0.8^2)} = 69.12, \tag{5.30}$$

respectively, as is indeed the case in Figure 5-5.

In summary, it is clear that if it turns out to be optimal to run a large phase II trial, the endpoint whose mean is most strongly correlated with θ_3 is likely to be an attractive choice. For investments above a certain threshold, it dominates against all biomarkers with lower correlation. There may however also be situations where it is found optimal to run a small phase II trial. A less strongly correlated, but cheaper, biomarker may then provide an efficient alternative.

5.4.3 Information and decision-making

We have discussed how to evaluate and choose between different ways of purchasing information that may be available from biomarkers. But purchasing information in optimal fashion is of little value if the information is not used in an adequate way. Let us focus on the information needed for the decision about whether to progress to phase III. One important question is whether a phase II trial is needed to make this decision. A related topic is discussed by Ades et al. (2004), who describe a general approach to deciding whether an experiment should be carried out. They use a quantity referred to as the expected value of sample information, which is the difference between the expected value of the optimal decision made after data have been collected, and the expected value of an optimal decision made immediately, without collecting new data. For the experiment to be carried out, the expected value of sample information must be larger than the cost of sampling. In our setting, such an approach could be used to decide whether it is worth to run a phase II trial with a certain investment, instead of moving directly to phase III.

As pointed out by Burman and Senn (2003), a decision point such as a phase III go/no go decision gives an option to discontinue a treatment that is not promising enough to justify further investment. As discussed in Section 5.3.5, the expected utility

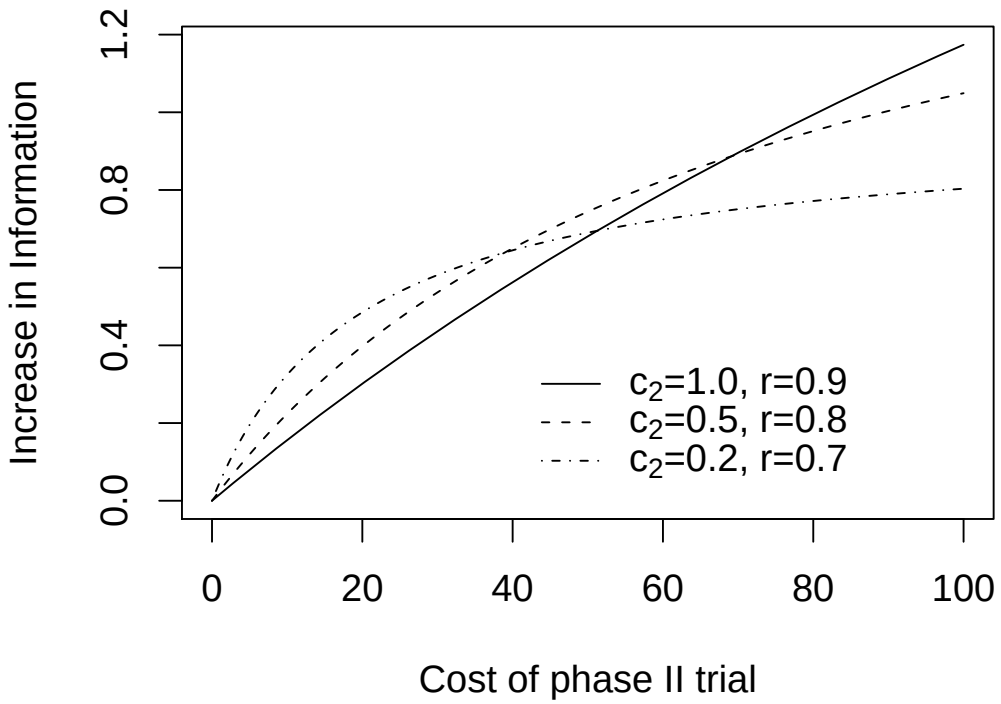

Figure 5-5: Information obtained for θ_3 depending on the choice of biomarker. All biomarkers have $t_1^2 = 0.04$ and sample variance $\sigma_2^2 = 1$.

is maximised if we continue to phase III whenever the expected utility of such a decision, conditional on $Z_2 = z_2$ and n_2, is positive. The value of the phase II trial is that it provides the option not to run the phase III trial if the expected utility after phase II turns out to be negative. If the expected utility after the phase II trial, conditional on $Z_2 = z_2$ and n_2, is positive with probability close to 1, the probability of running the phase III trial will also be close to 1. The phase II trial then has a negligible impact on the go/no go decision, and may turn out to be redundant.

These properties can help to explain the results seen in graphs such as Figure 5-4, where the expected utility typically has a minor local minimum in n_2. To be useful, the phase II trial has to be large enough to be able to sway the investigator's mind about running the phase III trial. If the probability of progressing to phase III does not change much regardless of the results, the phase II trial is of limited value. As there is also a cost per observation $c_2 > 0$ in phase II, this cost may sometimes outweigh the contribution that the phase II trial can make to the decision-making process.

Similar considerations can help to explain the discontinuity seen in Figure 5-2, where the choice of n_2 jumps from 0 to a positive value as the correlation r passes a certain threshold. Suppose that for a given decision problem with fixed r, it is optimal to make an investment $C_2^\star$ in the phase II trial. We know from (5.23) that the information obtained about θ_3 for a given phase II investment is increasing in r. We also know from Figure 5-2 that the optimal phase II sample size is monotonically increasing in r. When solving the decision problem for a smaller r, we will thus get a smaller phase II sample size, and due to smaller r and n_2, less information about θ_3. Eventually, as r continues to decrease, there comes a point when the increase in information about θ_3 is very small. The phase II trial is then unlikely to have any impact on the decision about whether to run the phase III trial. The value of the option described by Burman and Senn (2003) decreases, and the information that has been purchased no longer compensates for the cost of sampling. This results in the situation shown in the upper left panel of Figure 5-4, for $n_{3f} = 1046$ and $r = 0.45$. The figure shows that the expected utility at the local maximum, obtained for just under 60 observations in phase II, is smaller than the expected utility at $n_2 = 0$.

5.5 Joint optimisation of phase II and phase III

5.5.1 Adapting phase III sample size based on phase II results

We will now discuss the problem of optimising the phase III sample size for a given posterior distribution of θ_3 after phase II, as well as assumptions about g/c_3 and σ_3^2. This posterior distribution can also be viewed as the prior distribution of θ_3 before phase III. Suppose that the prior distribution of θ_3 before phase III is normally distributed according to

$$\theta_3 \sim N(\mu_2, \tau_2^2),$$

with $\mu_2 = 0$ and $\tau_2^2 = 0.04$. For now, we are not concerned with how this distribution has been derived, and simply take it as given. We will use it to illustrate the general method of optimising the phase III sample size, for a given prior distribution of θ_3 before phase III. To this end we consider three specific ratios for g/c_3 and set $\sigma_3^2 = 1$. As for the moment we are not concerned with the design of phase II, no assumptions are needed for c_2/c_3, r, or the prior distribution of θ_2. Pezeshk et al. (2009) have solved this problem for a more general gain function, so our situation with a constant g/c_3 can be viewed as a special case of their more general framework. A description of how to find the optimal phase III sample size, for a given prior distribution of θ_3 before phase III and assumptions about g/c_3 and σ_3^2, is given in Section 5.8.2.

The left panel of Figure 5-6 shows an example of how the assurance, as defined in equation (5.14), depends on the phase III sample size n_3. The centre panel shows

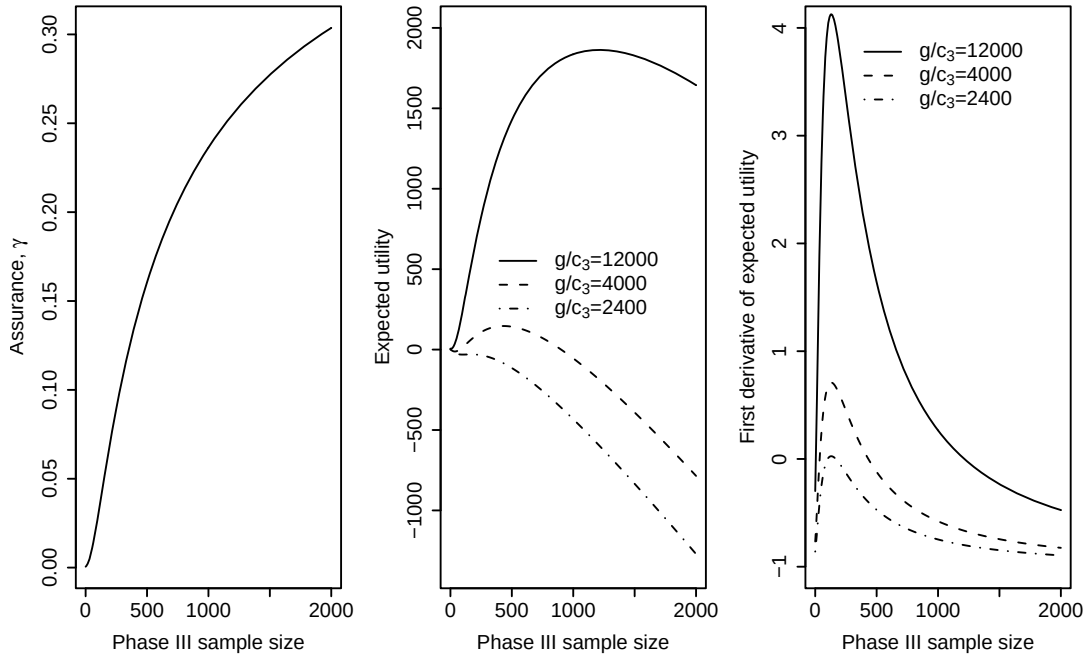

Figure 5-6: The left panel shows the assurance for different values of n_3. The centre panel shows the expected utility for different values of n_3 and different ratios g/c_3. The right panel shows the derivative of the expected utility with respect to the phase III sample size n_3. The prior distribution for θ_3 is assumed to be normal with mean zero and variance 0.04.

the expected utility as a function of n_3, for three specific choices of g/c_3 and $c_3 = 1$. For g/c_3=12000 and g/c_3=4000, the expected utility has a global maximum for our optimal phase III sample size $n_3^\star$. Perhaps more surprisingly, all three curves have a local minimum for small n_3. This can be explained mathematically by treating n_3 as continuous, and evaluating the first derivative of the expected utility with respect to n_3 numerically. The result is displayed in the right panel of Figure 5-6, where we see that each curve has two roots where the derivative equals zero. We would then expect each curve in the left panel of Figure 5-6 to have two stationary points, as is indeed the case.

The curve for $g/c_3 = 2400$ shows that the optimal phase III sample size may be obtained as $n_3 \to 0^+$. The expected utility of running such a phase III trial would be just above zero. We note that for $n_3 = 0$, the probability of rejecting the null hypothesis equals 0.0005, so the expected utility of running a phase III trial with $n_3 = 0$ equals $0.0005 \times 2400 = 1.2$. As the phase III sample size is increased the expected utility decreases and soon becomes negative. In practical applications we would however not choose to run a very small phase III trial for this reason of optimality. In fact, such an approach would be similar to running a very small study with no

treatment effect and hoping for a type I error. Hence, it may be useful to put a lower bound on the phase III sample size, for example $n_3 \geq n_{3,min}$, where $n_{3,min}$ is driven by the requirement to collect a minimal amount of safety data for the drug to be approved. Regulatory authorities frequently make such demands, to be able to make an appropriate benefit/risk assessment of the drug under investigation. The possibility of imposing a lower constraint on the maximal sample size will be further discussed in Section 5.5.2, where the choice of phase II sample size is also considered. Another alternative, which will not be considered in this chapter, could be to require that the assurance of the phase III trial must not be below a certain threshold. Such a constraint would avoid solutions of the type seen for the curve for $g/c_3 = 2400$, where $n_3^\star \to 0^+$.

5.5.2 Optimisation of both phase II and phase III sample sizes

Choice of phase III sample size

In Section 5.3, we discussed the design of phase II trials given a fixed phase III sample size n_3. Even though a posterior distribution $\pi_{\theta_3|Z_2}(\theta_3|z_2)$ is available after phase II, n_3 was not allowed to depend on phase II data. Rather than using the phase III designs described in Section 5.3, it is natural to ask if there is a way to make use of the additional information obtained in phase II when designing the phase III trial. The aim is to find the optimal phase III sample size, given the prior $\pi_{\theta_3|Z_2}(\theta_3|z_2)$ that has arisen from phase II.

Suppose that phase II is run with sample size n_2. After having observed the phase II results, the posterior distribution $\pi_{\theta_3|Z_2}(\theta_3|z_2)$ can be calculated according to (5.7), using the expressions for μ_2 in (5.8) and for τ_2^2 in (5.9). From Section 5.5.1, we have a method to find the optimal phase III sample size $n_3^\star$ for a given prior for θ_3 before phase III. This method can be used for any phase II result and sample size. The phase III sample size is then a function $n_3(Z_2, n_2)$ of the phase II sample size and results. We note that μ_2 and τ_2^2 are completely determined by $Z_2 = z_2$, n_2 and the prior distributions before phase II. We shall be writing $n_3(Z_2, n_2)$, even though figures will sometimes display n_3 as a function of μ_2 rather than Z_2.

After a phase II trial with sample size n_2 we can, given the joint prior distribution $\pi_{\theta_2,\theta_3}(\theta_2, \theta_3)$, $Z_2 = z_2$ and n_2, calculate μ_2, the posterior mean of θ_3 after phase II. Consider again our core example, with $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) before phase II according to (5.16) with $r = 0.8$. For selected values of the phase II sample size n_2, Figure 5-7 shows how to optimally choose the phase III sample size based on μ_2, the prior mean of θ_3 after phase III. The three curves for different n_2 have a similar shape and as we would expect, μ_2 is more important for determining the phase III sample size than n_2. To better understand the discontinuities seen for a certain value of μ_2, we can again consider Figure 5-6. The curves in Figure 5-6 that display the expected utility as a function of n_3 typically have two local maxima, one

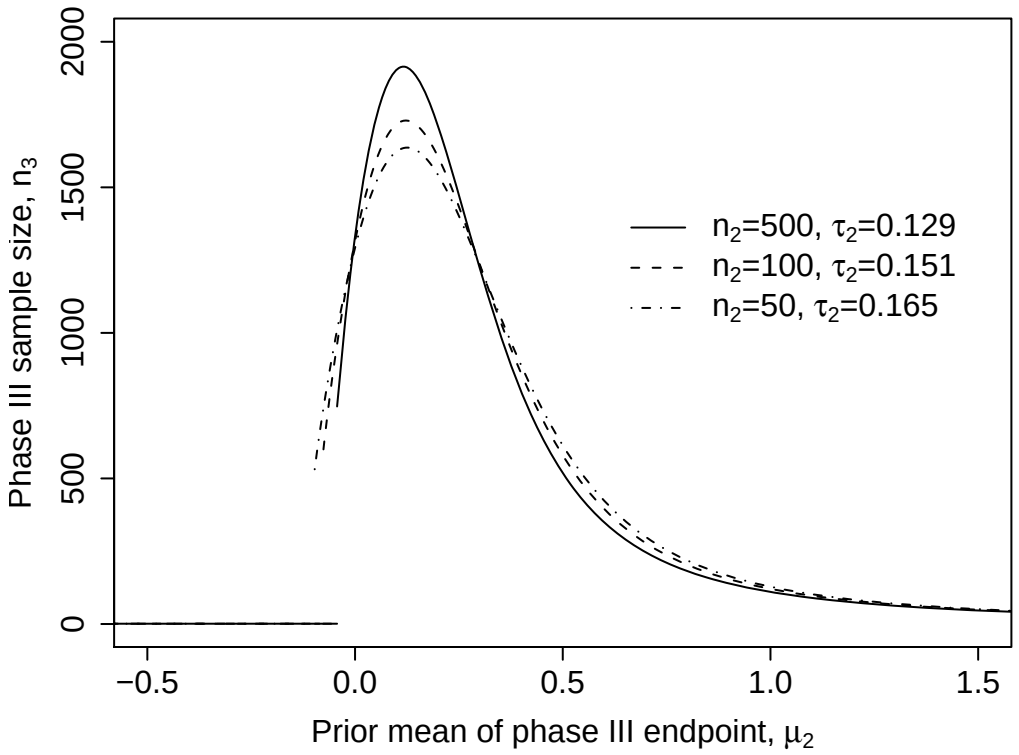

Figure 5-7: Optimal phase III sample size depending on prior mean of θ_3 before phase III, for different choices of phase II sample size n_2. The phase III sample size has been optimised for our core example, with $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16) with $r = 0.8$.

for $n_3 = 0$ and one for $n_3 > 0$. As we in Figure 5-7 move towards a more negative μ_2, there will for each curve come a point where the local maximum for $n_3 = 0$ is higher than the maximum for $n_3 > 0$. If there is no lower constraint on the phase III sample size the optimal solution will then be to set $n_3 = 0$, as was the case for the curve with $g/c_3 = 2400$ in Figure 5-6.

Choice of phase II sample size

We now have a method for how to choose phase III sample size, given a certain result $Z_2 = z_2$ in a phase II trial with n_2 observations per treatment group. The next step is to use this method to optimise the phase II sample size. To illustrate how the design of phase II changes when we allow optimisation of n_3 based on the results in phase II, we return to our core example in Section 5.3. To this end we assume a correlation $r = 0.8$, reflecting a situation where the phase II results can have substantial impact on the phase III sample size. Further details for how to solve this type of decision problem,

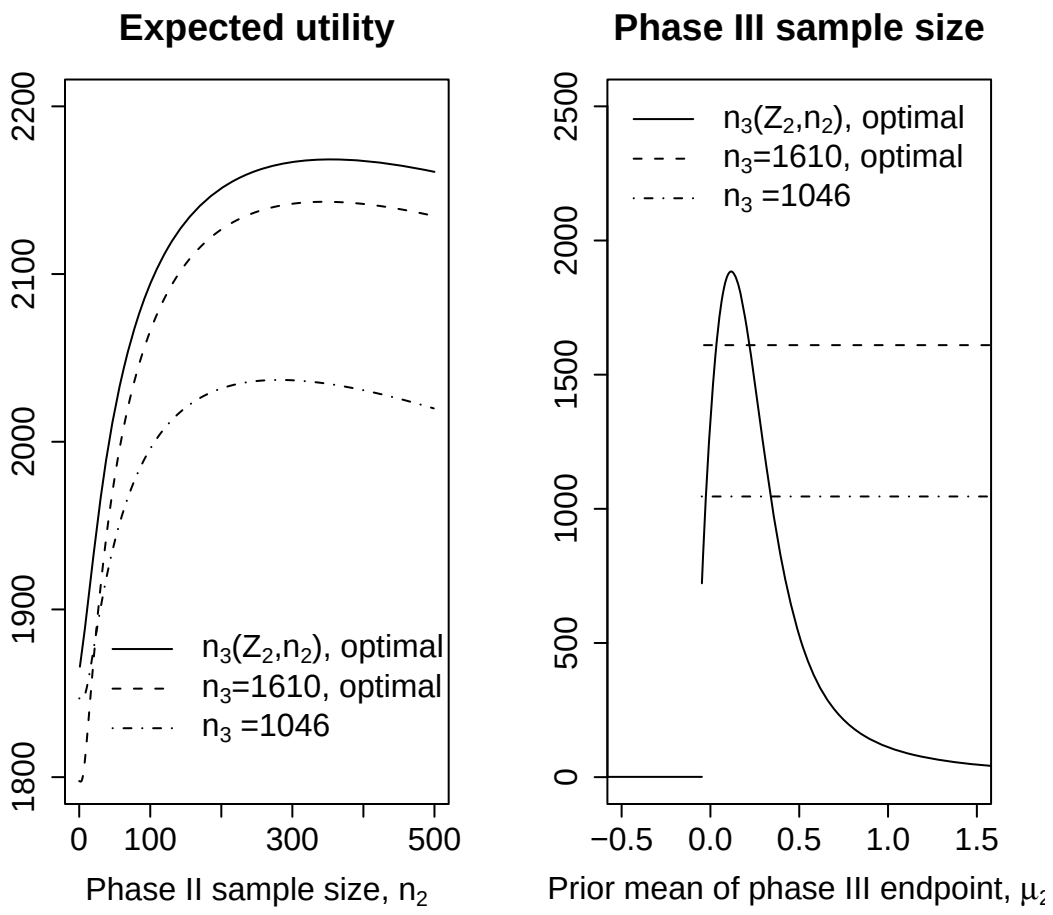

Figure 5-8: The left panel shows how the expected utility depends on the phase II sample size, for three different approaches to choosing phase III sample size. The right panel shows how the phase III sample size depends on the prior mean after phase II. The designs have been optimised for our core example, with $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16), with $r = 0.8$.

where we seek to find both the optimal phase II sample size $n_2^\star$ and an optimal rule for choice of phase III sample size based on the results in phase II, are provided in Section 5.8.2.

The expected utility for the optimal solution is displayed with a solid line in the left panel of Figure 5-8. The phase II trial now plays the role of guiding the phase III sample size, in addition to providing the go/no go decision for progress to phase III. We see that the expected utility is slightly higher than when n_3 was not allowed to depend on phase II data. Some improvement of the expected utility is then to be expected, as an optimal rule for choosing n_3 based on Z_2 is used. The right panel of Figure 5-8 shows the optimal rule for how to choose n_3 based on the results in phase II, for the case when the phase II sample size is optimally chosen. For comparison, the fixed sample

sizes $n_3 = 1046$ and $n_{3f}^\star = 1610$ are also displayed. The optimal phase III sample size varies with μ_2, the prior mean of θ_3 after phase II, and can depending on μ_2 be both higher or smaller than $n_{3f}^\star = 1610$.

It is also of interest to revisit the rule for how to choose phase II sample size based on the correlation r, now that n_3 is allowed to depend on phase II data. We saw in Figure 5-2 that there was a cut-off, so that for $r < r^\star$, the optimal phase II sample size $n_2^\star$ equals zero. Figure 5-9 shows how our core example changes when n_3 is allowed to depend on phase II data. The behaviour is similar, but there is also a key difference compared to when $n_3 = 1046$ or $n_{3f}^\star$. While there is still a cut-off $r^\star$ below which it is not worthwhile to run a phase II trial, there is no longer a discontinuity in $n_2^\star$ as a function of r. Likewise, we saw in the left panel of Figure 5-8 that the expected utility, displayed with a solid line, no longer has a local minimum for small n_2. This reflects the additional role of the phase II results, as they are now also used to find an appropriate phase III sample size. Both of these changes, compared to when n_3 was not allowed to depend on phase II data, are related to the discussion in Section 5.4 about the local minima and discontinuities seen for fixed phase III samples. Since phase II data now play a role in deciding the phase III sample size, even a small amount of information can make a difference to the decision. So for the purpose of fine-tuning the phase III sample size, the benefits of a small phase II trial can justify the costs. In contrast, a larger amount of information would be required to change the investigator's mind about whether to progress to phase III.

Lower bound on phase III sample size

When seeking the optimal phase III sample size for a given prior mean μ_2, we have so far not imposed a lower bound $n_{3,min}$ on the phase III sample size. In practice there is however likely to be such a constraint, as regulatory authorities typically require a minimum amount of safety data. Solutions with n_3 close to zero are theoretically interesting, but of little practical use in the drug development process. Furthermore, a trial with a very small number of patients is unlikely to be acceptable to the wider scientific community. Hence, we will now present some results for the case when n_3 has a lower bound. With a lower constraint for n_3, it is according to our model still possible to set $n_3 = 0$ and not run the phase III trial. It is however not possible to run a phase III trial with $n_3 = 0$, type I error probability $\alpha = 0.0005$ and constant power function equal to 0.0005 for all values of θ_3.

When solving the decision problem for the case when n_3 has a lower constraint, we can proceed in a similar away as when n_3 is allowed to take any value. Further details of how to solve the problem are provided in Section 5.8.2. Table 5.1 shows the expected utility that can be maintained for a given value of $n_{3,min}$. It is re-assuring to know that unless the constraint imposed is very high, the impact on the expected utility is

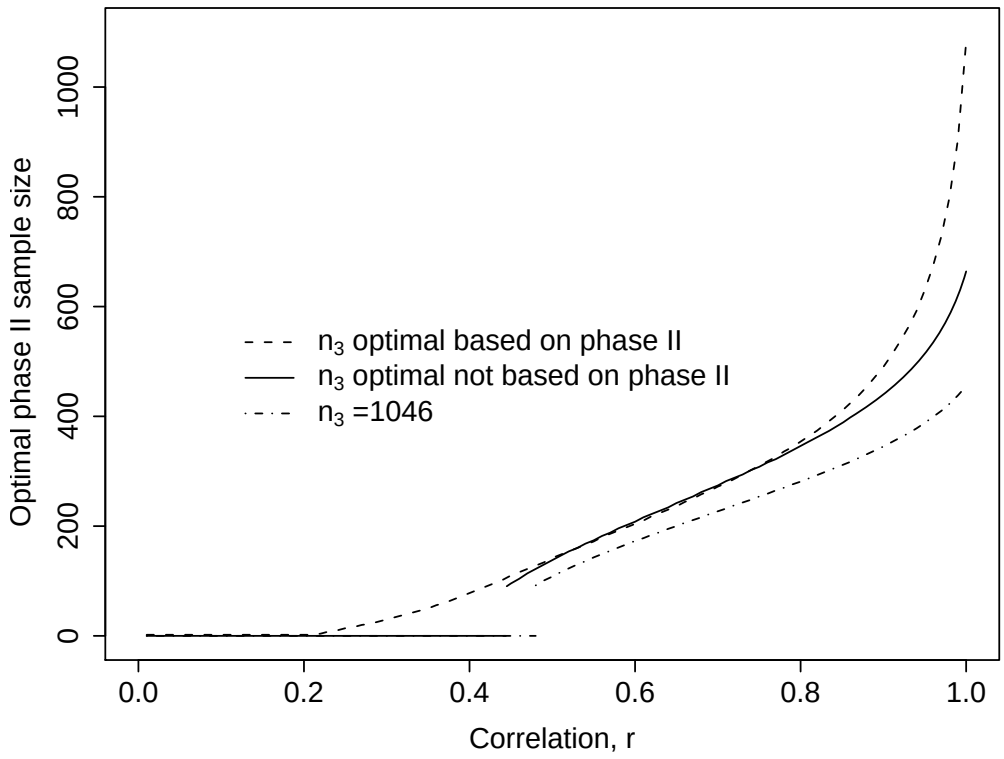

Figure 5-9: Phase II sample size depending on correlation r, for three different approaches to determining phase III sample size. The designs have been optimised for our core example, with $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16), with $r = 0.8$.

limited.

Figure 5-7 shows that the optimal phase III sample size varies with μ_2, and that μ_2 was more important for the choice of phase III sample size than n_2. Hence, it is somewhat surprising that the loss incurred from imposing a lower constraint on the phase III sample size is not larger than that shown in Table 5.1. One reason is that the very small phase III trials that are allowed when there is no lower constraint, with a low cost and probability of success, only have a small impact on the expected utility. Another reason has to do with the prior distribution of θ_3 before phase II, which is normal with mean $\mu_1 = 0$ and variance $\tau_1^2 = 0.04$. The phase II results, which in Figure 5-8 give an optimal phase III sample size that falls below a given constraint, may under this prior occur with only a small probability. This will in turn limit the impact on the expected utility, in particular for the cases in Table 5.1 where the lower bound is not very high.

$n_{3,min}$	$n_2^\star$	P(Progress to phase III)	$E(C_3\|$ Progress to phase III)	PoS| Progress to phase III	$E(U)$
0	354	1.00	967	0.267	2169
1	354	1.00	967	0.267	2168
200	354	0.63	1545	0.426	2167
1046	353	0.62	1568	0.431	2166
1610	359	0.59	1717	0.455	2149
2000	380	0.57	2000	0.486	2113
3000	441	0.51	3000	0.570	1877
4000	476	0.46	4000	0.636	1558

Table 5.1: Optimal phase II sample size, probability of progress to phase III, expected phase III investment C_3 given progress to phase III, probability of success (PoS) given progress to phase III and expected utility of designs with lower constraint $n_{3,min}$ on phase III sample size n_3. The designs have been optimised for our core example, with $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16), with $r = 0.8$.

Threshold for progress to phase III

Let us now revisit the problem of finding a decision rule for progress to phase III. After the phase II trial, we have

$$\theta_3 \sim N(\mu_2, \tau_2^2)$$

and as before, we proceed to phase III if $\mu_2 \geq \mu_2^\star$ and stop otherwise. In Section 5.3 we derived an analytical expression for $\mu_2^\star$, for the case when n_3 was fixed. Let us now consider how to derive the threshold $\mu_2^\star$ for the case when n_3 is allowed to depend on the phase II results. We must first know the phase III sample size that is optimal for a given phase II result $Z_2 = z_2$ and phase II sample size n_2, or equivalently for a certain prior mean μ_2 and prior variance τ_2^2 for θ_3 after phase II. From the results in Section 5.3, we know that $\mu_2^\star$ can be calculated as the solution to

$$\mu_2^\star = \frac{z_{1-\alpha} - \Phi^{-1}(1 - (n_3^\star(\mu_2^\star, \tau_2^2) \times c_3 + a_3)/g)\sqrt{1 + n_3^\star(\mu_2^\star, \tau_2^2)\tau_2^2/(2\sigma_3^2)}}{\sqrt{n_3^\star(\mu_2^\star, \tau_2^2)/(2\sigma_3^2)}}.$$

As there is not a simple analytical expression for how to calculate $n_3(\mu_2^\star, \tau_2^2)$, we have to find $\mu_2^\star$ numerically in this case. When searching for $\mu_2^\star$, it is helpful to know that conditional on τ_2^2, the expected utility is monotonically increasing in μ_2. For fixed n_3, it follows from the definition of assurance in (5.14) that this must be the case. We show in Section 5.8.2 that the expected utility, conditional on τ_2^2, is monotonically increasing in μ_2, also when n_3 is optimally chosen based on phase II data. To find the optimal go/no go decision, it then remains to perform a one-dimensional search for $\mu_2^\star$.

Even though $\mu_2^\star$ has to be found numerically, we can prove a simple property that relates to the analytical solution obtained when n_3 is not allowed to depend on phase II data. We know that there are now more options for how to choose n_3 than when n_3 is fixed, so the best phase III design must consequently be better or at least as good as when n_3 is fixed. Because of the increased flexibility for how to choose n_3, the posterior mean cut-off $\mu_2^\star$ must be smaller than when n_3 is fixed.

5.5.3 Group sequential phase III design

Introduction

So far we have assumed that phase III is run as a fixed sample trial with n_3 patients per treatment group, either chosen based on phase II results or according to some other criteria. In many areas of clinical research, for example in oncology and the cardiovascular area, phase III trials are often monitored using group sequential methodology. It is thus of interest to assess how the properties of the phase II and phase III designs change, if the phase III trial is instead made group sequential. As discussed in Chapter 4, regulatory authorities may question the practice of buying back type I error in one-sided group sequential tests. Consequently, the one-sided group sequential tests considered in this section will all have non-binding futility boundaries. We know from Chapter 4 that group sequential designs with non-binding futility boundaries have attained type I error rate $< \alpha$ and hence, are a little conservative. We refer to Chapter 4 for further details about such designs.

We know from Chapter 4 that ρ family error spending designs with non-binding futility boundaries are efficient and close to optimal. We shall therefore restrict our attention to this class of group sequential designs. The parameter ρ decides the amount of early stopping, with the probability of early stopping being monotonically decreasing in ρ. To give a substantial probability of early stopping, we set $\rho = 1$ for both the upper and the lower boundary. This gives boundaries that are broadly similar to those proposed by Pocock (1977). Compared to a more conservative boundary, O'Brien and Fleming (1979) say, our stopping rule gives a low expected sample size, at a price of a higher maximal sample size.

The number of analyses K has an important effect on the efficiency of a group sequential design. We know from publications such as Barber and Jennison (2002), that the biggest efficiency gains are achieved when moving from $K = 1$ to $K = 2$, while the efficiency gains achieved by increasing the number of groups beyond 5 are typically modest. Hence, we will consider the cases $K = 1$, 2, 3 and 5. Eales and Jennison (1992) found that while some additional efficiency could be achieved by optimising the group sizes of a K-group sequential tests, equally spaced analyses are rather efficient. Assuming K equally spaced analyses, the only parameter that has yet to be decided is the maximal sample size $n_{3,max}$. We shall be considering three approaches to deciding

the maximal sample size $n_{3,max}$ of group sequential designs:

1. Choosing $n_{3,max}$ to obtain power $1-\beta$ at the clinically relevant difference of $\theta_3 = \delta = 0.2$. To find $n_{3,max}$, we search for the maximal sample size needed to satisfy the power requirements. For $n_{fix} = 1046$, $\rho = 1$, $\alpha = 0.0005$ and $\beta = 0.1$, we have $R = 1.13$ and $n_{3,max} = 1183$ for $K = 2$, $R = 1.19$ and $n_{3,max} = 1243$ for $K = 3$ and $R = 1.22$ and $n_{3,max} = 1281$ for $K = 5$.

2. Choosing $n_{3,max} = n^{\star}_{3,max}$ to maximise the expected utility, without letting $n^{\star}_{3,max}$ depend on phase II data. The optimal maximal sample size of this $\rho = 1$ error spending design depends on the number of groups K.

3. Choosing $n_{3,max} = n^{\star}_{3,max}(Z_2, n_2)$ to maximise the expected utility, based on phase II data. The optimal maximal sample size of this $\rho = 1$ error spending design also depends on the number of groups K.

No other group sequential designs than the ones stated above will be considered. Given the results in Chapter 4, we don't expect that further optimisation of the group sequential boundary in phase III would have a significant effect. Setting $\rho = 1$ gives boundaries with substantial possibilities of early stopping and, compared to less aggressive boundaries with a higher ρ, rather high maximal sample sizes to get a certain power at a given effect size. It is appropriate to consider this type of boundaries in the absence of a threshold for $n_{3,max}$, above which it is difficult to recruit patients or handle other logistical aspects of the trial. If it is expected that patient recruitment will be difficult it may well be useful to consider group sequential boundaries that are derived for a higher value of ρ, which require a smaller maximal sample size to achieve a specific power.

Many of the qualitative results shown in previous sections apply also when phase III is group sequential, so we shall not repeat all of the sensitivity analyses performed for the case when phase III is a fixed sample trial. Rather, we have identified three qualitative features of our model that change when phase III becomes group sequential:

1. The phase II trial becomes less important, as the group sequential design in phase III can provide an efficient stopping rule based on θ_3. There are now several go/no go decisions at the interim analyses in phase III, which partly can replace the go/no go decision for proceeding to phase III based on phase II data.

2. The maximal sample size in a group sequential phase III trial is often considerably larger than the sample size of a fixed sample trial, as early stopping lowers the expected sample size and cost. We can aim for high power at small values of θ_3, as the early stopping makes it possible to avoid increases in expected sample size at high values of θ_3.

3. Adding additional interim analyses to a non-adaptively chosen phase III design increases the expected utility more than updating the choice of sample size based on the phase II results.

Illustrating these concepts will be the focus of this section about group sequential designs.

Optimising phase II sample size given group sequential phase III design

We shall now return to our core example to illustrate how some of the key features of the phase II and phase III designs change when the phase III trial is group sequential. For a given group sequential phase III design we can, in a similar fashion as in Section 5.3, perform a one-dimensional search for the phase II sample size that maximises the expected utility. Further discussion about how to solve the decision problem are provided in Section 5.8.2. In Figure 5-2, we illustrated how the proportion of resources spent in phase II depends on the correlation r and the cost ratio c_2/c_3. Figure 5-10 shows a similar illustration for the case when phase III is group sequential, where two different approaches have been used to choose $n_{3,max}$. The two panels to the left show designs with 90% power at $\theta_3 = 0.2$. In the two panels to the right, $n_{3,max}$ is chosen to maximise the expected utility, without being allowed to depend on phase II data. The results shown for $K = 1$ in Figure 5-10 are the same that were shown in Figure 5-2, when the ratio $c_2/c_3 = 0.2$. An important observation is that for fixed correlation r, the proportion of resources used in phase II is lower for $K = 2$ than for $K = 1$. This is likely to be because part of the role of the phase II trial, a go/no go decision rule, can now be played by the interim analyses in the phase III trial.

The point that phase II becomes less important when phase III is group sequential is illustrated even more clearly in the left panel of Figure 5-11, which shows the expected utility depending on phase II sample size, for designs with 90% power at $\theta_3 = 0.2$. We see that despite the rather high correlation of $r = 0.8$, the optimal phase II sample size $n_2^\star = 0$ for $K = 5$, and benefits of a phase II trial are very modest for $K = 2$. In the right panel it is shown that there is a stronger case for a phase II trial when the phase III sample size is larger, as a result of being chosen to maximise the expected utility. The benefit of phase II decreases with K, but also for $K = 5$ we obtain $n_2^\star > 0$. The value shown for $n_2 = 0$ refers to moving directly to phase III, without a phase II trial. We also note that just like the fixed sample designs, all the group sequential designs in Figure 5-11 have a minor local minimum for small n_2.

Things could be different, in both the left and right panel of Figure 5-11, if $c_2 << c_3$. Making phase III group sequential may then have less of an impact on the phase II sample size, which could then still be substantial. The observations collected in phase III on the other hand have the advantage that regardless of the cost ratio c_2/c_3, they can contribute to the final hypothesis test.

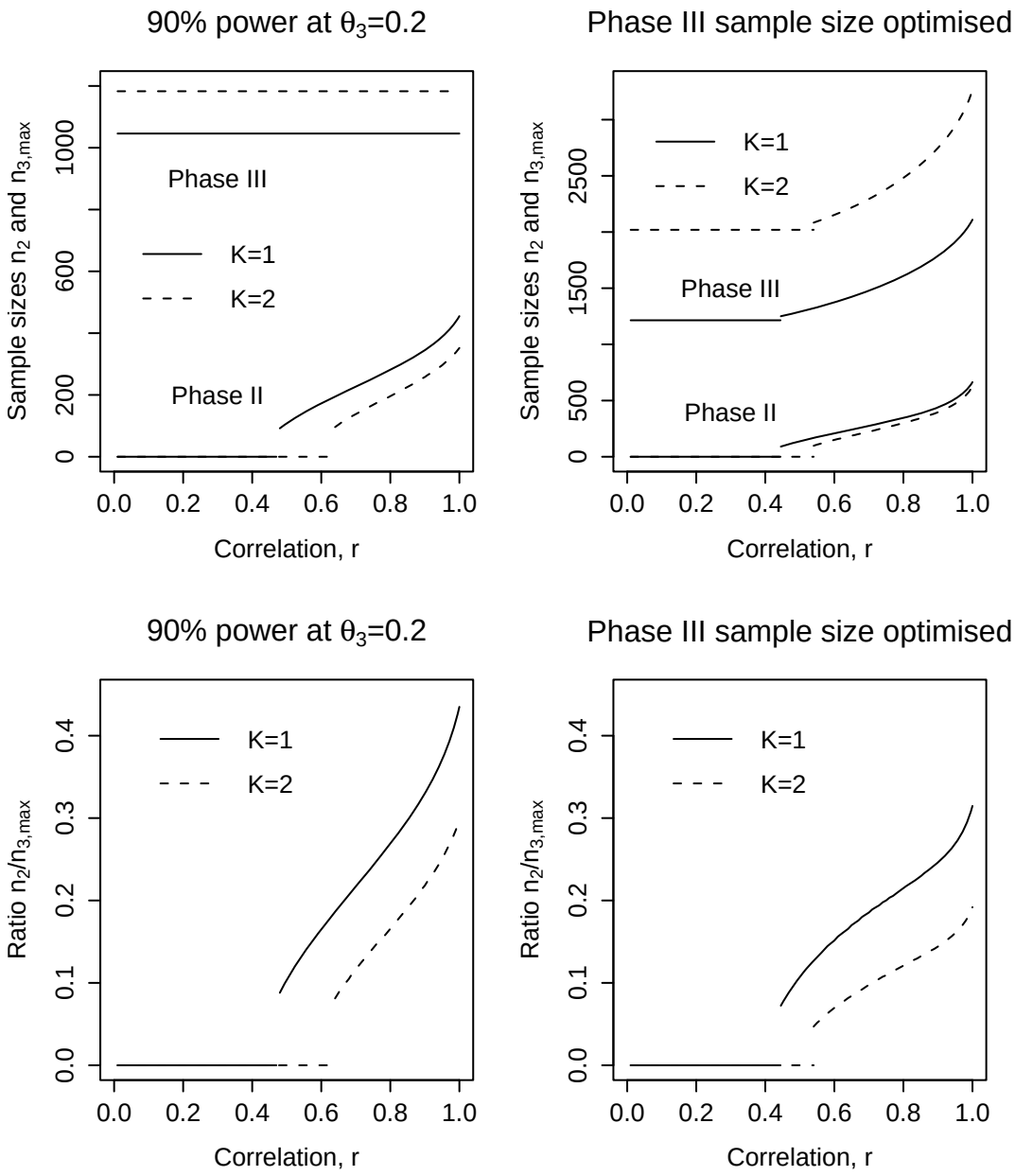

Figure 5-10: Optimal choices of phase II sample size, maximal phase III sample size and sample size ratio, for fixed sample phase III designs and $\rho = 1$, $K = 2$ error spending designs. The left panel shows results for when the phase III sample size is chosen to obtain 90% power at $\theta_3 = 0.2$. In the right panel, the phase III sample size is chosen to maximise the expected utility, without being allowed to depend on phase II data. Results for $K = 1$ are shown with solid line and for $K = 2$ with dashed line. The designs have been optimised for our core example, with $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16), with $r = 0.8$.

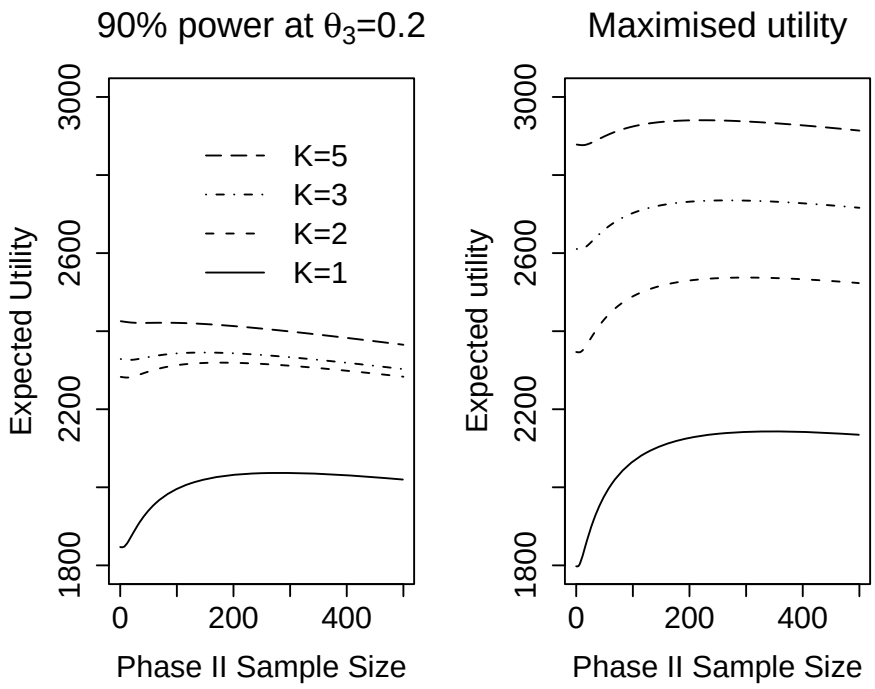

Figure 5-11: Expected utility versus phase II sample size, for two different approaches to choosing the maximal sample size $n_{3,max}$ of the phase III trial: The left panel shows designs with 90% power at $\theta_3 = 0.2$. The right panel shows designs with maximal sample size $n_{3,max}$ chosen to maximise the expected utility, when $n_{3,max}$ is not allowed to depend on phase II data. The K-group, $\rho = 1$ error spending designs have been optimised for our core example, i.e. for $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16), with $r = 0.8$.

Optimising the maximal sample size of group sequential phase III design based on phase II results

For a K-group sequential design, with equally spaced analyses say, we can now in a similar way to what was done for fixed sample trials in Section 5.5.1, search for the value of $n_{3,max}$ that is optimal, given n_2 and $Z_2 = z_2$, from a decision-analytic perspective. For our core example, with $r = 0.8$, Figure 5-12 shows how this maximal sample size $n_{3,max}(Z_2, n_2)$ depends on μ_2, the prior mean before phase III. Also displayed are the fixed sample sizes $n_3 = 1046$ and $n_3 = 1610$, where the former is chosen to achieve 90% power at $\theta_3 = \delta$ and the latter to maximise the expected utility. Finally, the optimal maximal sample sizes $n_{3,max}$, when the maximal sample sizes has not been chosen based on phase II data, are shown for $K = 2$, $K = 3$ and $K = 5$. It is noteworthy that for all the group sequential designs displayed in Figure 5-12, the optimal maximal sample size is highly dependent on the number of groups in the group sequential design. All the designs give good opportunity to stop with a fairly small sample size, if a boundary is

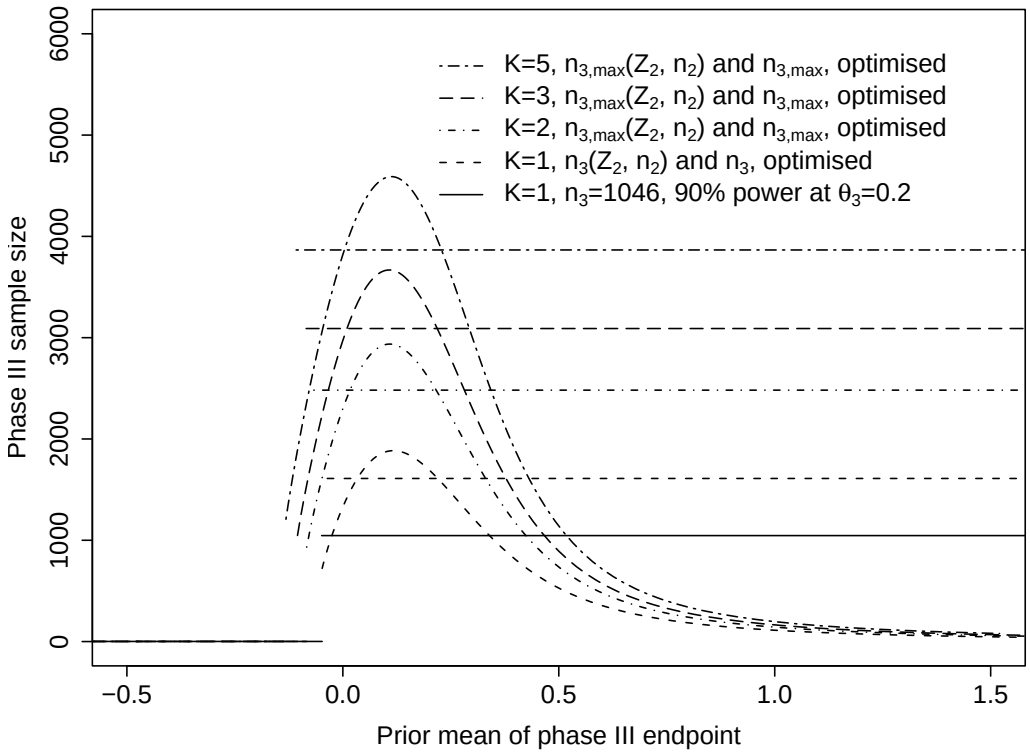

Figure 5-12: Phase III sample size depending on μ_2, for three fixed sample designs and maximal sample size for two group sequential designs. The designs have been optimised for our core example, with $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16), with $r = 0.8$.

crossed.

As the sample size increases, the probability of rejecting H_0 for small effect sizes decreases. The power for a given value of θ_3 depends on $\theta_3\sqrt{n_3}$ and there is, taking into account the inflation factor of the group sequential boundary of $R = 1.22$, roughly a threefold increase in $n_{3,max}$, from $n_3 = 1046$ for $K = 1$ to $n_{3,max} = 3866$ for $K = 5$. The optimal 5-group sequential design, when $n_{3,max}$ is not allowed to depend on phase III data, thus achieves a certain power at an effect size that is smaller by a factor of approximately $\sqrt{3}$, compared to the fixed sample phase III trial which achieves 90% power at $\theta_3 = 0.2$.

Optimising the phase II sample size and the maximal sample size of a group sequential phase III design

Given that we know how to choose the maximal sample size $n_{3,max}$ of the group sequential design for a given prior, we can use the approach from Section 5.5.2 to find

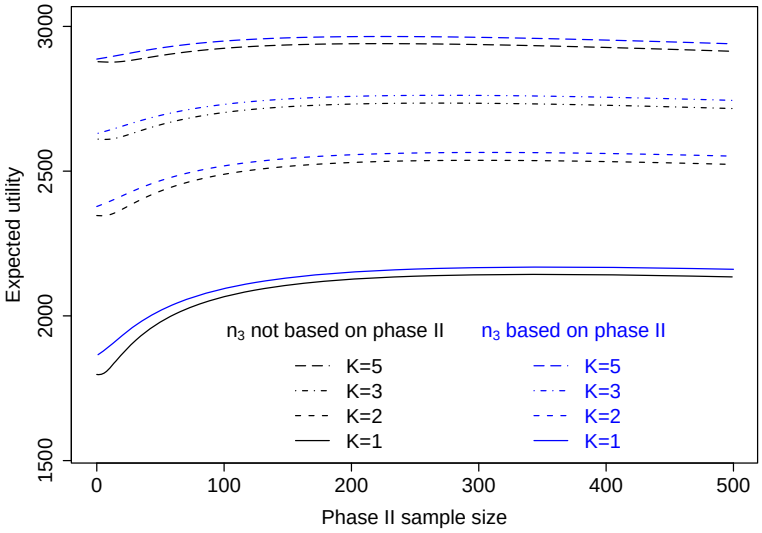

Figure 5-13: Expected utility versus phase II sample size n_2, for two different approaches to choosing the maximal sample size $n_{3,max}$ of the phase III trial: For the black lines, $n_{3,max}$ is not allowed to depend on phase II data, while for the blue lines, $n_{3,max}$ is chosen based on phase II data. The K-group, $\rho = 1$ error spending designs have been optimised for our core example, with $g/c_3 = 12000$, $c_2/c_3 = 0.2$ and prior distribution for (θ_2, θ_3) according to (5.16), with $r = 0.8$.

the optimal phase II sample size $n_2^\star$, as well as the rule $n_{3,max}(Z_2, n_2)$ for how to choose the maximal sample size of the group sequential design in phase III. We have done this for our core example, with $r = 0.8$. We see in Figure 5-13, where the black lines show the same results as the right panel in Figure 5-11, that the phase II sample size becomes less important to the expected utility as we increase K, the number of groups in the group sequential design. The optimal phase II sample decreases with K for the results displayed with blue lines in Figure 5-13, from over 350 observations for $K = 1$ to slightly below 230 observations for $K = 5$. The trend is similar when the optimal phase II sample size is decreased from just under 350 observations for $K = 1$ to about 220 observations for $K = 5$. There is however some scope for increasing the expected utility by running a phase II trial also when phase III is group sequential, in particular when phase II is also used to guide the phase III sample size. For $K = 5$, the expected utility can for example be increased by about 3%, compared to when moving directly to phase III without a phase II trial. Adding interim analyses to the phase III design appears to be much more important than updating the phase III sample size based on phase II results. Adding one interim analysis to the fixed sample design (black solid line) increases the expected utility by about 18%. This can be compared with an improvement of just over 1% that is achieved by running a fixed sample trial, where the

sample size is allowed to depend on phase II data. The expected utility is improved by 37% for the $K = 5$ design compared to the $K = 1$ design. An improvement of 38% is possible if in addition the maximal sample size of the group sequential design is chosen based on the phase II results. These results are broadly in agreement with what we found earlier in this thesis, about choosing future group sizes based on the observed treatment effect within a single study. For group sequential designs for simultaneous testing of superiority and non-inferiority, non-adaptive 3-group sequential designs were found to be more efficient than 2-group adaptive designs. In the present problem it is worth remembering that it should be a drawback that the adaptation is based on a biomarker for θ_3, rather than the phase III endpoint.

5.6 A numerical example

We will now illustrate how our model can be used with a numerical example. We are interested in using our model to assess the potential benefits of first running a smaller phase II study, before embarking on a phase III trial. We shall let the phase III sample size be guided by the phase II results. At first, a fixed sample phase III trial will be assumed. To achieve the benefits of early stopping we will thereafter be considering group sequential designs, with one interim analysis before the final analysis.

Gittins and Pezeshk (2000b) cite a clinical trial that was carried out by a UK-based pharmaceutical company, where the drug concerned was already on the market. We will use some of the numerical values in their example to illustrate how our model could be applied. The aim of the clinical trial was to broaden the indication for the drug in question and thus increase the sales of the product. To assess the feasibility of running such a trial, one low estimate and one high estimate for future annual sales had been established. It was assumed that the new broader indication would result in between 5% (low estimate) and 50% (high estimate) increase in annual profit. The gain obtained in case of regulatory approval, £15 million for the low estimate and £150 million for the high estimate, were obtained by discounting the resulting cash flow of the future annual sales. The cost per patient in the phase III trial was estimated to be $c_3 =$ £4000, which in our notatation implies $g/c_3 = 3750$ for the low sales estimate and $g/c_3 = 37500$ for the high sales estimate. Gittins and Pezeshk (2000b) comment that a more careful analysis would include a start-up cost, a_3 in our notation, for the phase III trial, in addition to the cost per patient. We shall start by setting $a_3 = 0$, but will later in this section get back to the implications of introducing a start-up cost in phase III.

Before any trials on the broader indication, the prior beliefs about θ_3 could be summarized by the prior distribution

$$\theta_3 \sim N(\mu_1, \tau_1^2)$$

with $\mu_1 = 0.41$ and $\tau_1^2 = \mu_1^2/4$. This prior distribution has most of its probability mass at positive values of θ_3, and is more optimistic than the prior distribution centred at zero considered in the previous sections. Moreover, $\sigma_3^2 = 2$ was thought to be a reasonable assumption for the common within-group sample variance of the phase III endpoint. We shall also be making the slightly simplified assumption that the ratios $g/c_3 = 3750$ and $g/c_3 = 37500$ are achieved whenever the null hypothesis is rejected at the $\alpha = 0.0005$ level. In the model of Gittins and Pezeshk (2000b), g depends also on z_3, in such a way that the full gain functions that we have specified are obtained only if z_3 is sufficiently high. We show in Section 5.8 how our decision problem can be approached for this more sophisticated gain function.

Once the posterior distribution for θ_3, as well as assumptions for g/c_3 and σ_3^2, are available, it is straightforward to use our model to decide how to design the phase III trial. We now make some additional assumptions that were not specified in the problem solved by Gittins and Pezeshk (2000b). We follow the approach in Table 5.1 and impose a lower bound, in this case 100 patients per treatment group, on the phase III sample size. This is a very sensible feature, as it has a negligible impact on the expected utility, but avoids designs with phase III sample size close to zero.

As before, a biomarker is used in phase II to obtain information about θ_3. Suppose that two different biomarkers are available for use in the phase II study, both with sample variance $\sigma_2^2 = 1$ but with different costs per observation, $c_{21} = 0.1c_3$ for biomarker 1 and $c_{22} = 0.5c_3$ for biomarker 2. The means of the two biomarkers have the same prior variance $t_1^2 = 0.25$, but different correlations with θ_3, $r_1 = 0.8$ and $r_2 = 0.9$, respectively. Let us now consider the possibility of using either of these biomarkers in a pre-study, preceding our phase III trial. We know from Section 5.4, that since the biomarkers have the same prior variance t_1^2 and sample variance σ_2^2, the two biomarkers will deliver the same amount of information about θ_3, for a given investment C_2, if

$$C_2^\star = \frac{2\sigma_2^2(c_{22}r_1^2 - c_{21}r_2^2)}{t_1^2(r_2^2 - r_1^2)} = \frac{2(0.5c_3 \times 0.8^2 - 0.1c_3 \times 0.9^2)}{0.25(0.9^2 - 0.8^2)} = 11.2c_3 = £44800. \quad (5.31)$$

Figure 5-14 shows how the expected utility depends on the phase II investment and the choice of biomarker. We see from Figure 5-14 that the curves for the expected utility cross at $C_2^\star = £44800$ for both the low and high sales estimate, which is in agreement with the result in (5.31). This condition about the efficiency of the biomarkers applies regardless of the phase III sample size and which sales estimate is assumed to be the most appropriate. We also see that the expected utility is considerably larger if the phase II investment is increased to a level $C_2 > C_2^\star$. Biomarker 1 would be the better option if the phase II investment were constrained according to $C_2 \leq C_2^\star$. If this is not the case, biomarker 2 is to be preferred.

We see in Figure 5-14 that for $g/c_3 = 37500$, the optimal phase II investment is

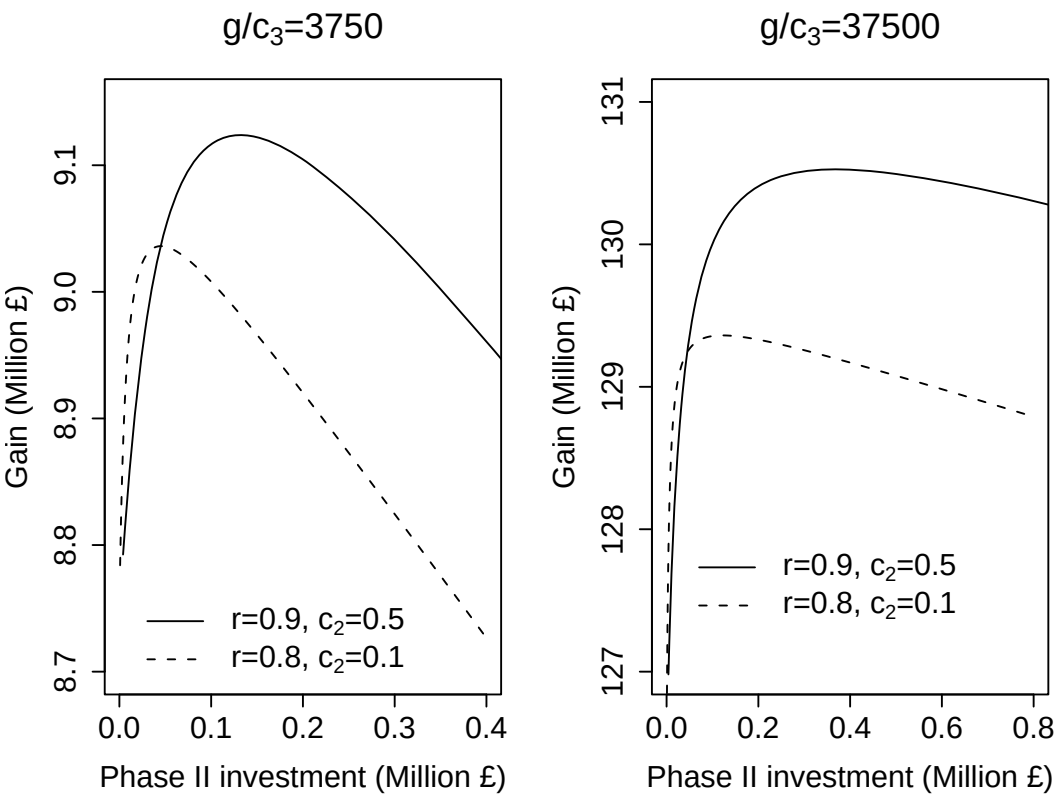

Figure 5-14: Expected utility for different choices of phase II sample size n_2 and two different biomarkers, when n_2 is used to guide the phase III sample size. The decision problem has been solved assuming that the prior distribution for θ_3 is normal with mean 0.41 and variance $0.41^2/4$, a prior variance for θ_2 of 0.25, $\sigma_2^2 = 1$ and $\sigma_3^2 = 2$.

considerably higher than when $g/c_3 = 3750$. When $g/c_3 = 3750$, about 4%, or £0.3 million pounds, can be gained by running a pre-study, compared to moving directly to a fixed sample phase III trial. For $g/c_3 = 37500$, about 3%, or £4 million, can be gained.

It has been shown in previous sections that substantial benefits can be achieved by making phase III group sequential. We therefore consider running a $\rho = 1$ error spending design with two equally spaced analyses. We do not expect the choice of biomarker to change in the group sequential setting, as biomarker 1 is more efficient than biomarker 2 only for very small investment levels. Hence, we will focus on biomarker 2 for the remainder of this section. We will now focus our attention on running a phase II trial with this biomarker, and assess the impact on the choice of phase III sample size, as well as the go/no go decision.

The two upper panels of Figure 5-15 show how the phase III sample size depends on the phase II results, for the case when biomarker 2 is used in phase II. The results for a fixed sample phase III trial are shown with a solid line, while the results for a

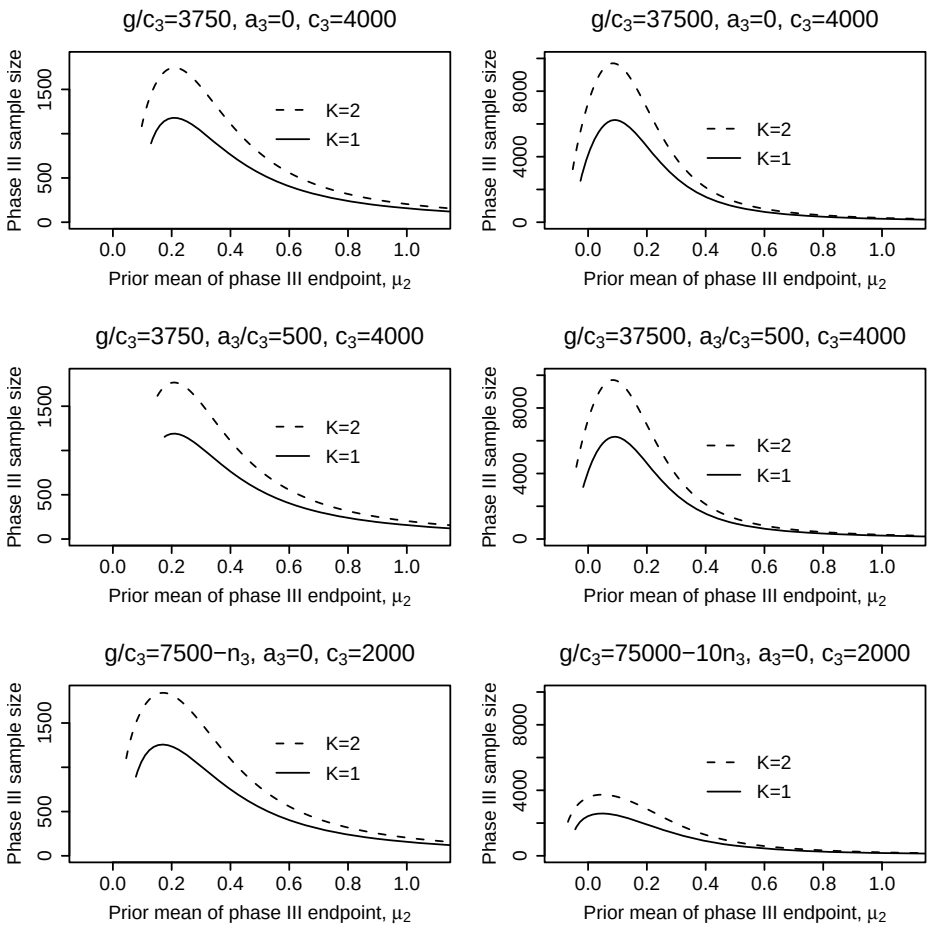

Figure 5-15: Phase III sample size for fixed sample trial (solid line) and maximal phase III sample size for $K = 2$, $\rho = 1$ error spending design (dashed line), depending on the prior mean of θ_3 before phase III. The six subplots show results for different assumptions about gain function, cost per patient and start-up cost. The decision problem has been solved assuming a prior distribution for $\theta_3 \sim N(0.41, 0.41^2/4)$, a prior variance for θ_2 of 0.25, $r = 0.9$, $c_2/c_3 = 0.5$, $\sigma_2^2 = 1$ and $\sigma_3^2 = 2$.

group sequential design are shown with a dashed line. The maximal sample size is increased by a higher amount than what a typical inflation factor for an error spending design with $K = 2$ analyses would motivate. Hence, the group sequential design aims at achieving a higher power at a given effect size. The cut-off $\mu_2^\star$, which defines the go/no go decision rule, is somewhat higher for a fixed sample trial than when phase III is group sequential. For $g/c_3 = 3750$, we have $\mu_2^\star = 0.13$ for $K = 1$ and $\mu_2^\star = 0.10$ for $K = 2$. We would expect this behaviour, as the possibility of early stopping in the group sequential design makes it feasible to run a phase III trial with a lower expected cost. We also note that for both $K = 1$ and $K = 2$, the cut-off is considerably higher for $g/c_3 = 3750$, than for $g/c_3 = 37500$. This property of the go/no go decision rule is consistent with what we observed in previous sections about the threshold for progress

to phase III. It reflects the increased confidence in the compound that is required to pursue a phase III trial, when the forecasts for future sales are less optimistic.

Figure 5-15 also shows results for two sensitivity analyses that we shall now consider. The first is related to the question of including a start-up cost $a_3 > 0$ in the model for the cost structure phase III trial. So far it has been assumed that $a_3 = 0$, letting the cost of the phase III trial be proportional to the number of patients. In the group sequential setting, it is however very realistic to think that there is some additional start-up cost that is not directly proportional to the number of patients at the time of the interim analysis. The two panels in the centre of Figure 5-15 show that including a start-up cost $a_3/c_3 = 500$, rather than $a_3 = 0$, has a notable impact on the decision rule for progress to phase III. The start-up cost will sometimes mean that the expected utility of running phase III becomes negative, when it otherwise would have been positive. The impact is more pronounced for the low sales estimate, when the start-up cost is a higher proportion of the total gain g. The cut-offs for μ_2 previously cited, 0.13 for $K = 1$ and 0.10 for $K = 2$, are now shifted to 0.18 and 0.15, respectively. If the expected utility of running the phase III trial is positive, the optimal phase III sample size is however very similar to when $a_3 = 0$.

The two panels in the centre of Figure 5-16 show how the expected utility is affected by the start-up cost. As there was some change in how to choose n_3 based on μ_2, we would expect the start-up to have an impact on the properties of our model, in particular when $g/c_3 = 3750$ and the start-up costs eliminates a substantial proportion of the potential gain. We see in Figure 5-16 that the relative benefit of phase II trial is increased, from 4% to 7%, when a start-up cost for phase III is included for the lower sales estimate. Like in previous examples, it is also noteworthy that the group sequential phase III design delivers important efficiency gains compared to the fixed sample phase III trial. Overall, the impact of the start-up cost on the expected utility is less substantial than the impact on the decision rule for progress to phase III.

An important feature of this example is that the prior mean is set two standard deviations of the prior distribution for θ_3 above zero. Based on the prior information it is thus very probable that the drug has a positive treatment effect compared to control. It is thus to be expected that the relative increase in expected utility, about $3 - 4\%$, is rather modest, as the go/no go decision rule should be less uncertain. It is however worth remembering that a modest increase on the relative scale may correspond to an important difference in absolute numbers, especially if the sponsor has difficulties to fund the trial.

Our final sensitivity analysis concerns the situation where the gain function depends on the number of patients accrued in the phase III trial. This is a very reasonable assumption, as the number of patients in phase III has an impact on the duration of the trial and time of patent life left when the drug is approved. For the low sales

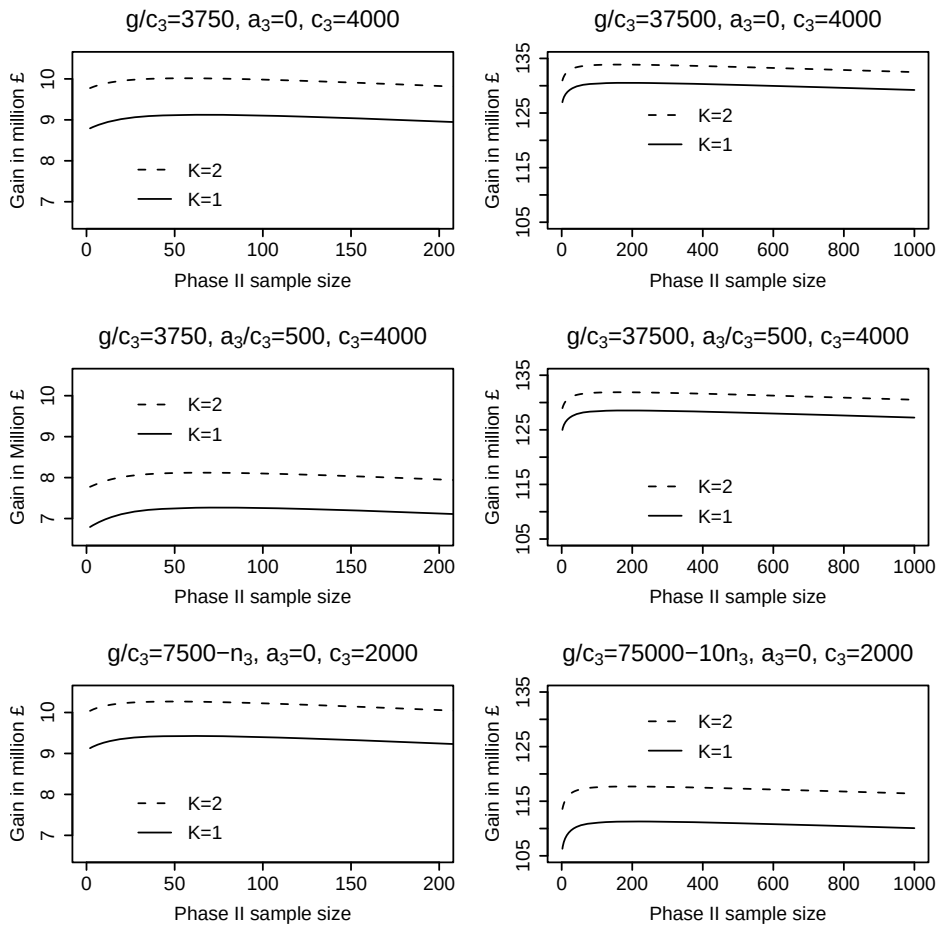

Figure 5-16: Expected utility for different choices of phase II sample size n_2, when n_2 is used to guide the phase III sample size. Results are shown for a fixed sample trial phase III trial (solid line) and for $K = 2$, $\rho = 1$ error spending design (dashed line). The six subplots show results for different assumptions about gain function, cost per patient and start-up cost. The decision problem has been solved assuming a prior distribution for $\theta_3 \sim N(0.41, 0.41^2/4)$, a prior variance for θ_2 of 0.25, $r = 0.9$, $c_2/c_3 = 0.5$, $\sigma_2^2 = 1$ and $\sigma_3^2 = 2$.

estimate, we shall consider the case $g(n_3) = \max((7500 - n_3)c_3, 0)$, where n_3 is the number of subjects on termination, while the cost $c_3 =$ £2000. Compared to our previous model, the cost per patient in phase III is reduced from £4000 to £2000. The potential gain if the drug is approved is however also decreased, by £2000 per patient. For the high sales estimate, the cost per patient remains $c_3 =$ £2000, while $g/c_3 = \max((75000 - 10n_3)c_3, 0)$, as we find it reasonable to assume that the time component will be proportional to the potential gain that will be achieved if the new indication of the drug is approved. In this sensitivity analysis, the gain function is zero if the sample size at termination $n_3 \geq 7500$, for both the low and high sales estimate, so it would not make sense to schedule interim analyses for $n_3 \geq 7500$. Defining the gain function in this manner thus gives a definitive upper constraint for the maximal

sample size. We see in the bottom right panel of Figure 5-15, that for both $K = 1$ and $K = 2$, the phase III sample size is decreased considerably when this model for g is used for the high sales estimate. The situation is different for the low sales estimate, where the cost structure for the phase III observations has become more favourable to the trial sponsor. To see why this is the case, we note that £2000$(1 + I(z_3 \geq z_{1-\alpha}))$ is now charged for having one patient per treatment group, instead of $c_3 =$ £4000 regardless of the result in the phase III trial.

The two lower panels in Figure 5-16 show how the expected utility depends on the phase III sample size, when the gain function depends on n_3. We see a largely similar pattern as before, even though things differ slightly between the low and high sales estimate. In the former case, the new cost structure makes it possible to increase the expected utility, as the cost per patient in phase III is decreased from £4000 to £2000. In the latter case, the dependence of the gain function on n_3 means that the expected utility is lower, even though c_3 is decreased by a half. It is not surprising that the results for the high and low sales estimate vary, as the high sales estimate assumes a tenfold increase compared to the low sales estimate. A different approach would have been to assign a prior probability to these sales estimates and work with the expectation under this prior. A third possibility would be to let the gain function depend explicitly on θ_3 and z_3, so that the different sales forecasts are built in to the definition of g.

5.7 Discussion

We have seen that the correlation between θ_2, the mean of the biomarker, and θ_3, the mean of the clinical endpoint, as well as the cost of sampling, are crucial when deciding the phase II sample size. The prior uncertainty of θ_2 and θ_3 is also important. We have derived an expression for the information obtained for a certain investment, which can be used to assess if a biomarker is appropriate to use in the phase II trial. Suppose that it turns out to be optimal to make a large investment in phase II. A biomarker that is expensive but strongly correlated with the mean of the phase III endpoint is then likely to be more useful than a cheaper biomarker that is less strongly correlated. If only a small investment is made in phase II, a cheaper biomarker, which still has a substantial correlation, may provide a useful option.

Our model can be used to find the optimal phase II sample size, and hence, the amount of resources spent in phase II. We have found that the resources spent in phase II increase when the phase III investment is high, and there is uncertainty about whether the drug will progress to phase III. The phase II trial can however become redundant, if for example the ratio g/c_3 is so high that the phase III trial will be carried out with probability close to one. The possibility to stop a project is important, in

particular when there is no group sequential monitoring in phase III. This can, in combination with the fact that the information in phase II may be cheaper than in phase III, justify an investment in phase II even though the observations in phase II are not used in the the final hypothesis test in phase III.

In the most general version of our model, n_3 is allowed to depend on phase II data. The importance of this adaptation can be assessed by comparing with the situation when n_3 is optimised, but not allowed to depend on the phase II data. The efficiency gains of the first approach compared to the latter are rather modest, but we would still recommend making use of the information gathered in phase II. In practice, it may be useful to have a working assumption for the phase III sample size before the phase II results are obtained. If desirable, the phase III sample size can then be updated once the phase II results are available. We would recommend to use a lower bound for the phase III sample size, as it was shown in Table 5.1 that the efficiency loss from such an approach is small. Using a very small phase III sample size, because of promising results for a biomarker in phase II, may on the other hand be a risky strategy that may undermine the credibility of trial results.

When a group sequential design is used in phase III, the expected utility can be improved by increasing the number of groups K. Another result of the possibility of early stopping is that the optimal maximal sample size of the group sequential design increases with K. We have found that increasing K, for example from $K = 1$ to $K = 2$, has a more pronounced impact on the expected utility than adaptively choosing the phase III sample size based on phase II data. In Section 5.5.3 an improvement of 18 % was observed when moving from $K = 1$ to $K = 2$, while running a fixed sample phase III trial with sample size updated based on phase II data gave an improvement of just 1%. The most efficient design is obtained when a group sequential design is used in phase III and the maximal sample size is chosen based on the phase II results.

It is not surprising that several aspects of the model change when a group sequential design is used in phase III. When phase III is group sequential, there is a possibility of early stopping. Hence, the early stopping in the phase III trial can partly replace the role that the phase II trial has in the go/no go decision. When the phase III sample size is chosen based on the phase II results, there is, provided that the cost and correlation structure of the biomarker is favourable, some use of phase II even when phase III is group sequential. This benefit does however decrease with the number of interim analyses K. It may however not be entirely realistic that the expected sample size reflects the entire cost of a group sequential phase III trial. In practice it may be more reasonable to include a start-up cost, which cannot be recovered through early stopping. We have seen in Section 5.6 that when this is the case, the importance of the phase II trial increases slightly.

In Section 5.6, we also experimented with introducing a gain function that was

decreasing in the phase III sample size n_3. It is then implicitly assumed that an increased number of patients will increase the duration of the trial, with a shorter remaining patent life once the drug is approved. Results were largely similar, but typically with a slightly smaller optimal phase III sample size. An even more careful analysis would make the gain function depend also on the phase II sample size n_2. It would be straightforward to extend our model to let g depend on n_2. As we are assuming that the biomarker in phase II can be rapidly observed, it would appear to be a reasonable approximation to let g be independent of n_2. The approximation will be particularly appropriate if there is not a big increase in recruitment time when increasing the phase II sample size. Moreover, it may be the case that even though the duration of the phase II trial is far from insignificant, other activities that are run simultaneously mean that the duration of the phase II does not in isolation decide when phase III can start. For the phase III trial that is the final activity before submission for regulatory approval, it is more obvious that an increase in the number of observations will delay a potential launch of the drug.

It is noteworthy that also for the rather modest ratios g/c_3 that we have considered, it is often worthwhile to move forward to phase III even if the posterior distribution of the treatment effect after phase II is not that impressive. There may be several ways to extend the model to make the decision rule for progress to phase III more conservative. Firstly, we have only required a statistically significant result, $p < 0.0005$ one-sided, for regulatory approval. In practice many phase III trials fail due to other issues such as unexpected safety problems. So if the assurance before embarking on a phase III trial according to our model is γ, the actual probability of regulatory approval is almost certain to be less than γ. A very simple model for safety could be that there is a risk of unexpected toxicity in phase III, that is independent of the efficacy of the drug. In such a situation, discounting the assurance by a factor to account for issues other than efficacy that may lead to the drug not being approved, would lead to a gain function $g \times \mathrm{P}(\mathrm{Safe})$ instead of g. Such an approach is unlikely to change the qualitative behaviour of our model.

A second extension could be to account for the fact that a pharmaceutical company can only run a limited number of late stage drug projects at the same time. A similar issue has been studied by Stallard (2003), who considered different drug projects competing for the same limited resource. The competition for resources between different projects may mean that some projects with positive expected utility may have to be parked if others are more promising. On the other hand, it could be argued that in such situations it might be possible to temporarily increase the volume of projects that can be run simultaneously, for example through cooperation with external contract research organisations.

A third extension could be to consider other prior distributions than our choice of

bivariate normal prior distribution. Even for a drug with a not so impressive prior mean, there is due to the tails of the normal distribution some probability mass at values of the treatment effect that make a phase III trial attractive. Perhaps a different choice of prior, for example with a point mass of zero treatment effect compared to control, would be an interesting extension to the model. We have nevertheless shown that phase II can be useful, in particular in identifying which treatments should be discontinued after phase II and increasing the probability of success conditional on progressing to phase III.

5.8 Derivation and implementation

5.8.1 Model derivation

Derivation of posterior distribution of θ_3 after phase II

In this section we will provide further details about the methods used to derive phase II and phase III designs that maximise the expected utility. We will be using the model assumptions in equations (5.3) and (5.11) as well as the fact that the property

$$\pi_{\theta_3|\theta_2,Z_2}(\theta_3|\theta_2,z_2)=\pi_{\theta_3|\theta_2}(\theta_3|\theta_2) \tag{5.32}$$

follows immediately from (5.3). Let us now use (5.32) to derive the posterior distribution of $\theta_3|Z_2=z_2$. We have

$$\pi_{\theta_3|Z_2}(\theta_3|z_2) \;=\; \int_{-\infty}^{\infty} d\theta_2 \pi_{\theta_3|\theta_2,Z_2}(\theta_3|\theta_2,z_2)\pi_{\theta_2|Z_2}(\theta_2|z_2),$$

and using the fact that θ_3 is conditionally independent of Z_2 given θ_2, we obtain

$$\pi_{\theta_3|Z_2}(\theta_3|z_2) \;=\; \int_{-\infty}^{\infty} d\theta_2 \pi_{\theta_3|\theta_2}(\theta_3|\theta_2)\pi_{\theta_2|Z_2}(\theta_2|z_2). \tag{5.33}$$

Recall from Section 5.2.3 that the conditional distribution of θ_3 given θ_2 is normal with mean $C+D\theta_2$ and variance τ_ϵ^2, while the posterior distribution of θ_2 given $Z_2=z_2$ is normal with mean m_2 and variance t_2^2. After θ_2 has been integrated out, it follows that the posterior distribution of $\theta_3|Z_2=z_2$ is normal with mean μ_2 and variance τ_2^2, where $\mu_2=C+Dm_2$ and $\tau_2^2=\tau_\epsilon^2+D^2t_2^2$.

Expected utility to be maximised

We will now derive the simplified expression for the expected utility, shown in (5.15). Taking the expectation of U as defined in (5.12), the expected utility in (5.13) can,

when g is allowed to depend on θ_3, z_3 and n_3, be written as

$$\begin{aligned}
E(U) &= -a_2 1_{n_2>0} - c_2 n_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dz_2 dz_3 d\theta_2 d\theta_3 f_{Z_2,Z_3,\theta_2,\theta_3}(z_2,z_3,\theta_2,\theta_3) \\
&\quad \times I(z_2 > z_2^\star)\{-a_3 - n_3(z_2,n_2)c_3 + I(z_3 > z_{1-\alpha})g(\theta_3,z_3,n_3(z_2,n_2))\} \\
&= -a_2 1_{n_2>0} - c_2 n_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dz_2 dz_3 d\theta_3 d\theta_2 \pi_{\theta_3|\theta_2}(\theta_3|\theta_2)\pi_{\theta_2}(\theta_2) f_{Z_2|\theta_2,\theta_3}(z_2|\theta_2,\theta_3) \\
&\quad \times f_{Z_3|\theta_3,\theta_2,Z_2}(z_3|\theta_3,\theta_2,z_2) I(z_2 > z_2^\star)\{-a_3 - n_3(z_2,n_2)c_3 + I(z_3 > z_{1-\alpha})g(\theta_3,z_3,n_3(z_2,n_2))\} \\
&= -a_2 1_{n_2>0} - c_2 n_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dz_2 dz_3 d\theta_3 d\theta_2 \pi_{\theta_3|\theta_2}(\theta_3|\theta_2)\pi_{\theta_2}(\theta_2) f_{Z_2|\theta_2}(z_2|\theta_2) \\
&\quad \times f_{Z_3|\theta_3,Z_2}(z_3|\theta_3,z_2) I(z_2 > z_2^\star)\{-a_3 - n_3(z_2,n_2)c_3 + I(z_3 > z_{1-\alpha})g(\theta_3,z_3,n_3(z_2,n_2))\},
\end{aligned} \tag{5.34}$$

where the last equality follows from the modelling assumptions in (5.3) and (5.11). We can combine (5.32) and (5.33) with the identity

$$\pi_{\theta_2}(\theta_2) f_{Z_2|\theta_2}(z_2|\theta_2) = \pi_{\theta_2|Z_2}(\theta_2|z_2) f_{Z_2}(z_2),$$

to obtain

$$\begin{aligned}
\int_{-\infty}^{\infty} d\theta_2 \pi_{\theta_3|\theta_2}(\theta_3|\theta_2)\pi_{\theta_2}(\theta_2) f_{Z_2|\theta_2}(z_2|\theta_2) &= \int_{-\infty}^{\infty} d\theta_2 \pi_{\theta_3|\theta_2,Z_2}(\theta_3|\theta_2,z_2)\pi_{\theta_2|Z_2}(\theta_2|z_2) f_{Z_2}(z_2) \\
&= \pi_{\theta_3|Z_2}(\theta_3|z_2) f_{Z_2}(z_2),
\end{aligned}$$

which inserted into (5.34) gives

$$\begin{aligned}
E(U) &= -a_2 1_{n_2>0} - c_2 n_2 + \int_{-\infty}^{\infty} dz_2 f_{Z_2}(z_2) I(z_2 > z_2^\star) \int_{-\infty}^{\infty} dz_3 \int_{-\infty}^{\infty} d\theta_3 \pi_{\theta_3|Z_2}(\theta_3|z_2) \\
&\quad \times f_{Z_3|\theta_3,Z_2}(z_3|\theta_3,z_2)\{-a_3 - c_3 n_3(z_2,n_2) + I(z_3 > z_{1-\alpha})g(\theta_3,z_3,n_3(z_2,n_2))\} \\
&= -a_2 1_{n_2>0} - c_2 n_2 + \int_{z_2^\star}^{\infty} dz_2 f_{Z_2}(z_2)\{-a_3 - c_3 n_3(z_2,n_2) + \int_{z_{1-\alpha}}^{\infty} dz_3 \\
&\quad \times \int_{-\infty}^{\infty} d\theta_3 \pi_{\theta_3|Z_2}(\theta_3|z_2) f_{Z_3|\theta_3,Z_2}(z_3|\theta_3,z_2) g(\theta_3,z_3,n_3(z_2,n_2))\}.
\end{aligned} \tag{5.35}$$

Here, $\pi_{\theta_3|Z_2}(\theta_3|z_2)$ denotes the pdf of the posterior distribution of θ_3, after having observed $Z_2 = z_2$ in a phase II trial with n_2 patients per treatment group. When $g = g(n_3(z_2, n_2))$ and does not depend on z_3 and θ_3, (5.35) reduces to

$$\begin{aligned} E(U) &= -a_2 1_{n_2>0} - c_2 n_2 + \int_{z_2^\star}^{\infty} dz_2 f_{Z_2}(z_2) \\ &\quad \times \{\gamma(z_2, n_2, n_3(z_2, n_2)) g(n_3(z_2, n_2)) - a_3 - c_3 n_3(z_2, n_2)\}. \end{aligned}$$

If g is independent also of n_3, which is assumed for most of this chapter, the formula displayed in equation (5.15) is obtained. Suppose that g does depend on n_3 and z_3, and is linear in θ_3,

$$g(\theta_3, z_3, n_3) = g_1(z_3, n_3(z_2, n_2)) \; + \; g_2(z_3, n_3(z_2, n_2)) \; \theta_3,$$

say. The expected utility can then be written as

$$\begin{aligned} E(U) &= -a_2 1_{n_2>0} - c_2 n_2 + \int_{z_2^\star}^{\infty} dz_2 f_{Z_2}(z_2) \\ &\quad \times \{-a_3 - c_3 n_3(z_2, n_2) + \int_{z_{1-\alpha}}^{\infty} dz_3 f_{Z_3|Z_2}(z_3|z_2) \\ &\quad \times (g_1(z_3, n_3(z_2, n_2)) + g_2(z_3, n_3(z_2, n_2)) \mu_3(z_3, n_3(z_2, n_2)))\}, \end{aligned}$$

where $f_{Z_3|Z_2}(z_3|z_2) = \int_{-\infty}^{\infty} d\theta_3 \pi_{\theta_3|Z_2}(\theta_3|z_2) f_{Z_3|\theta_3,Z_2}(z_3|\theta_3, z_2)$ and $\mu_3(z_3, n_3(z_2, n_2))$ is the posterior mean of θ_3 after phase III. This type of gain function bears similarities with the public health benefit function used by Gittins and Pezeshk (2000b).

We have focused our derivation on a very general case, where the gain function is not a constant and n_3 can be chosen based on the phase II results. The situation when we simplify to having g and n_3 as constants can be viewed as a special case of our derivation, which applies in these cases too.

The situation is very similar when, like in Section 5.5.3, phase III is group sequential. Instead of calculating the probability of exceeding $z_{1-\alpha}$ under the prior distribution of θ_3, the probability of stopping and continuing at each interim analysis has to be taken into account. When the gain function depends on n_3, it is important to note that the gain function takes different values depending on at which interim analysis the null hypothesis is rejected. The other important difference is that when the sample size depends on the amount of early stopping, we need to take this into account when calculating the cost per patient. For a group sequential phase III design, we are only charged c_3 times the expected sample size under the posterior distribution for θ_3 after phase II, as opposed to $c_3 \times n_3$ for a fixed sample phase III trial.

5.8.2 Implementation

Optimisation of phase III sample size

Sample size of phase III trial without interim analyses

In Section 5.5.1 we consider the problem of finding the optimal phase III sample size $n_3^\star$, given a prior distribution for θ_3 and assumptions about g/c_3 and σ_3^2. We will now provide further details of how to solve this problem. Suppose that

$$\theta_3|Z_2 = z_2 \sim N(\mu_2, \tau_2^2).$$

This is equivalent to having observed $Z_2 = z_2$ in a phase II trial with n_2 patients per group, for some z_2, n_2 and prior distribution. For given z_2 and n_2, we can calculate the expected utility of running a fixed sample phase III study with maximal sample size n_3. The next step is to perform a numerical search for the value $n_3^\star(Z_2, n_2)$ that gives the optimal phase III design for given $Z_2 = z_2$ and n_2. For a fixed sample phase III trial with sample size chosen based on the phase II results, $n_3^\star$ and the type I error probability $\alpha = 0.0005$ define the design of the phase III trial. If we instead are interested in $n_{3f}^\star$, which is not allowed to depend on the phase II results, we perform a two-dimensional search for the pair of sample sizes n_2, n_3 that maximise the expected utility.

Maximal sample size when phase III trial is group sequential

When, like in Section 5.5.3, a group sequential design is used in phase III, we have by choosing a ρ family error spending design already determined the shape of both the upper and lower boundary. If we are aiming for 90% power at $\theta = \delta$, it is straightforward to derive the maximal sample size $n_{3,max}$, through a one-dimensional numerical search. If the maximal sample size is chosen based on decision-analytic considerations, the problem is one-dimensional. If $n_{3,max}$ is allowed to depend on the phase II results, we can find the maximal sample size in a similar way as described above for a fixed sample trial. If we instead are interested in the fixed value of $n_{3,max}^\star$ that does not depend on the phase II results, we search for the pair of sample sizes $n_2, n_{3,max}^\star$ that maximise the expected utility.

Decision rule for progression to phase III

We saw in Section 5.3 that when the phase III sample size is not allowed to depend on phase II data, the cut-off $\mu_2^\star$ for progress to phase III can be found analytically. As explained in Section 5.5.1, this is not the case when the phase III sample size is chosen based on phase II data. When n_3 is allowed to vary depending on μ_2, we can still find $\mu_2^\star$ through a one-dimensional search, provided that the expected utility, for fixed τ_2^2, is monotonically increasing in μ_2. We show this in two steps:

1. We first make use of the fact that in all these designs, the expected utility must, for fixed τ_2^2 and n_3, be monotonically increasing in μ_2, the prior mean of θ_3 after phase II. This follows from the definition of assurance in equation (5.14).

2. To see what happens when n_3 is allowed to vary, consider two prior means, with $\mu_{21} < \mu_{22}$, and suppose that $n_3^\star(\mu_{21}, \tau_2^2) = n_{31}$. Using the same sample size n_{31} when the prior mean equals μ_{22} must for $\mu_{22} > \mu_{21}$ and fixed τ_2^2, according to the definition in (5.14), give a higher assurance and consequently higher expected utility. If $n_3^\star(\mu_{22}, \tau_2^2) \neq n_{31}$, this must be because another choice of phase III sample size gives an even higher expected utility, conditional on μ_{22} and τ_2^2, than n_{31}.

As the expected return is monotonically increasing in μ_2, a go/no go decision rule can be calculated by searching for $\mu_2^\star$, the value of μ_2 that gives an optimal phase III design with expected utility equal to zero.

Finding the optimal phase II sample size

Once we know the optimal phase III sample size $n_3^\star(z_2, n_2)$, for a given combination of $Z_2 = z_2$ and n_2, the next step is to find the phase II sample size $n_2^\star$ that maximises the expected utility. Given the results of a phase II design with n_2 observations, we can calculate $\pi_{\theta_3|Z_2}(\theta_3|z_2)$. With this posterior distribution available, we have a rule for how to choose the optimal phase III design. For any combination of $n_2, Z_2 = z_2$, the optimal phase III design can thus be regarded as known, regardless of whether phase III is a fixed sample design or group sequential. For some combinations $n_2, Z_2 = z_2$, the optimal design may be not to proceed to phase III and stop development of the drug in question.

To assess the optimal choice of n_2, we can integrate over the marginal distribution of Z_2 and obtain the expected utility as defined in (5.15), for different choices of n_2. Methods for numerical integration that are based on Simpson's rule are suitable in these situations, and are further described in Jennison and Turnbull (2000, Chapter 19). Provided that we know how to calculate the integral numerically, it remains to perform a one-dimensional search for $n_2^\star$, the value of n_2 that maximises (5.15).

CHAPTER 6

Discussion and conclusions

6.1 A broader drug development perspective

An important objective in drug development is to get drugs approved in areas where there is an unmet medical need. There are many areas where there is no satisfactory treatment available, or where the current standard treatment could be improved upon. There are consequently several players who have an interest in getting new, effective medicines approved. Patients and regulatory authorities are affected if fewer effective medicines are available, and the pharmaceutical industry is facing challenges such as expiring patents of existing drugs, increased costs of running clinical trials and difficulties with getting new drugs approved. There is a growing demand for novel, innovative approaches to how new medicines can be developed. Adaptive design of clinical trials is one new area that has received considerable attention. While it would be unrealistic to expect this class of designs to provide a solution to all the issues mentioned in this paragraph, it is hoped that it can be helpful in improving important aspects of clinical trial design.

This thesis deals with sequential and adaptive methods for clinical trials, and how such methods can be used to achieve efficient clinical trial designs. The efficiency gains that can be achieved through non-adaptive group sequential methods are well established, while the newer adaptive methods seek to combine the best of the group sequential framework with an approach that gives more flexibility. Within the context of a clinical development programme, both classes of designs can also be helpful in identifying which treatments should be progressed for further testing.

The increased costs of drug development, described by DiMasi et al. (2003), emphasise the need for designing development programmes that are more efficient. More efficient development programmes would make it possible to test larger number of promising candidate drugs. Adequately designed clinical trials can also be helpful in

identifying which treatments should be progressed to the later, more costly, stages of development. Hence, the efficient design of group sequential and trials can contribute to improving the drug development process and making more effective drugs available to patients. The four problems that have been studied in this thesis all relate to how clinical trials can me made more efficient and flexible, while maintaining the credibility of trial results. We will now briefly discuss the main results.

6.2 Summary of results

6.2.1 Optimal group sequential designs for simultaneous testing of superiority and non-inferiority

Confirmatory clinical trials, comparing the efficacy of a new treatment with an active control, typically aim at demonstrating either superiority or non-inferiority. We consider non-adaptive and adaptive group sequential designs that combine the two objectives. The difference between the two classes is that in adaptive group sequential designs, the future group sizes may be chosen based on the observed treatment effect.

For both classes of designs, we derive group sequential designs meeting error probability constraints which minimise the expected sample size, averaged over a set of values of the treatment effect. These optimised designs provide an efficient means of reducing expected sample size under a range of treatment effects, even when the separate objectives of proving superiority and non-inferiority would require quite different fixed sample sizes.

We also present error spending versions of the non-adaptive sequential designs, which are easily implementable and can handle unpredictable group sizes or information levels. The adaptive choice of group size yield some modest efficiency gains. Further reduction can however be achieved by adding another interim analysis to a non-adaptive group sequential design.

6.2.2 Control of type I error when applying the CRP principle in an error spending design

Error spending tests are efficient and can cope with unpredictable information sequences with exact control of the type I error rate. Since the conditional rejection probability (CRP) principle gives flexibility beyond what is available in error spending designs, it would be very beneficial to be able to combine the two approaches. We have investigated how to apply the CRP principle in a clinical trial with unpredictable information levels and present numerical examples of how the type I error can be inflated. A method that ensures protection of the type I error, by using a pre-specified combination rule, is discussed. It is found that such methods can lead to unequal

weighting of equally informative observations and thus some loss of efficiency. If it is very important to be able to adapt, it may well be that the increased flexibility that is provided compensates for the efficiency loss, which is often small. The conclusion from Chapter 3 is that care is needed when applying the CRP principle. The type I error can be inflated when applying the CRP principle in an error spending design, unless the conditional type I error is well defined in all situations when a re-design can occur.

6.2.3 Group sequential designs with non-binding futility boundaries

The efficiency gains that can be achieved through one-sided group sequential tests are well documented (Barber and Jennison, 2002; Jennison and Turnbull, 2006a). In Chapter 4 we address a problem that arises whenever there is a futility boundary in a one-sided group sequential test, but has sometimes been overlooked in the statistical literature. A new method to derive optimal group sequential designs with non-binding futility boundaries is presented and the results are compared against currently available designs with non-binding futility boundaries. The new optimisation method extends the method of dynamic programming, used in Chapter 2.

Using tests with non-binding futility boundaries should be re-assuring to regulators. They can be confident that the type I error is controlled, even if the futility boundary is not always applied. From the point of view of the trial sponsor, the boundaries are a little conservative, with an attained type I error that is smaller than α and a small loss of power as a result. By making comparisons with optimal group sequential designs with binding futility boundaries, we can quantify the efficiency loss incurred by the additional requirement that futility boundaries must be non-binding. Group sequential designs with non-binding futility boundaries nevertheless deliver substantial efficiency gains compared to fixed sample designs. It is found that both error spending designs and futility stopping based on predictive power give efficiency close to that of the optimal designs. The comparisons with the optimal group sequential designs means that we have two methods that are easy to use, with close to optimal efficiency.

6.2.4 Joint planning of phase II and phase III

In Chapter 5 we move beyond the individual trial, to consider the joint planning of one phase II trial and one phase III trial. We present a method to find the phase II and phase III sample sizes, as well as a decision rule for progress to phase III, that maximise the expected utility. We consider both the situation when the phase III sample size is fixed, and when it is optimally chosen based on phase II data.

In our model, with different endpoints in phase II and phase III, the utility of the phase II trial varies depending on the cost of sampling and the correlation with the mean of the phase III endpoint. Given a certain prior distribution, cost and correlation, we derive an expression for the utility of using a certain biomarker in phase III. This

expression can be used to choose between different biomarkers, with different cost and correlation structures.

The impact of running an error spending design, of the type used in Chapter 4 with a non-binding futility boundary, in phase III is also considered. It is found that running a group sequential design in phase III reduces the importance of the phase II trial. Each interim analysis of the group sequential design can be thought of as a new go/no go decision. Hence, the go/no go decision between phase II and phase III becomes less important.

We also assess what efficiency gains can be achieved, by adaptively choosing phase III sample size based on phase II data. To this end, a simplified version of the model, where the phase III sample size is not allowed to depend on phase II data, is compared to an optimal adaptation rule. It is found that adding an interim analyses to the phase III trial is more important to the expected utility than adaptively choosing the phase III sample size based on phase II data.

6.3 Discussion

6.3.1 Adapting future sample size based on observed data

Our results show that adaptive methods can provide some additional efficiency compared to classical group sequential designs, as well as increased possibilities to respond to new internal and external information. Care is however needed when applying adaptive methods. The conduct of clinical trials is important, and the logistical challenges of implementing adaptive methods can be considerable. Efficient non-adaptive group sequential designs are often easier to implement in practice, and have in the cases we have considered been quite competitive in terms of efficiency. The efficiency of non-adaptive and adaptive group sequential tests for a two-decision problem has been compared by Tsiatis and Mehta (2003), Brannath et al. (2006) and Jennison and Turnbull (2006a), among others. Our results for the three-decision problem in Chapter 2 are largely in agreement with their findings.

The problem of adaptively choosing the second group size, considered in Chapter 2, bears similarities with adaptively choosing phase III sample size based on phase II data, considered in Chapter 5. In both cases we find that the benefits of adaptation are modest, while greater efficiency can be achieved by increasing the number of groups in the non-adaptive group sequential designs. The adaptive group sequential designs in Chapter 2 are optimal from a theoretical perspective, but would in practice be difficult to implement for an unpredictable information sequence. We would prefer a pre-planned, non-adaptive group sequential design, using the error spending approach described in Section 2.2.5 to handle unpredictable information sequences. The situation is different in Chapter 5, where no adjustment to the final analysis is necessary because

of the adaptation based on phase II. Hence, we recommend the approach illustrated in Section 5.6, using the knowledge gained in phase II to update the phase III sample size.

6.3.2 Optimisation as an approach to clinical trial design

Throughout this thesis, decision analysis has been used as a device to derive clinical trial designs that maximise the expected utility, or equivalently, minimise the excepted cost. It is however our experience that in phase III in particular, it is still more common to design clinical trials based on the frequentist hypothesis test framework, where the sample size is chosen to achieve a certain power at a given effect size. Solving very general decision problems, involving many parameters, is difficult, and this could be one reason for why in practice, the approach is not so often applied. To be able to solve a decision problem, it is often necessary to make simplifying assumptions. On a less technical level, decision analysis is however not so different from what we do to make decisions in daily life. Different pros and cons are weighed against each other and the most difficult part may be to define the utility function. Even if the utility function is suitably defined, it may be difficult to specify some of the parameter values that are needed to solve the decision problem. This should however not be an excuse to base the design on ad hoc assumptions that are difficult to justify. We would recommend to first make an informed, conscious choice about which model and parameter values to use. Thereafter, it is important to perform robustness checks to assess how the uncertainty about the parameters in the model affects the decision-making.

The choice of prior distribution is another difficulty that should not be underestimated. O'Hagan et al. (2006) propose eliciting prior distributions based on the opinion of experts, but this possibility has not been considered in this thesis. It could be possible to use elicitation to determine the prior distribution in Chapter 5. For the problems in Chapter 2 and Chapter 4, the prior distribution is to a large degree decided by the need to control the error probability constraints at certain values of the treatment effect. It could be argued that the choice of prior distribution is more subjective in the problem in Chapter 5, where it more directly reflects the prior belief about the mean of the phase II and phase III endpoints.

6.3.3 Decision analysis in this thesis

Let us now consider how decision analysis has been applied in Chapters 2, 4 and 5 and assess what the differences and similarities are. In both Chapter 2 and Chapter 4, decision analysis is used to solve the constrained optimisation problem of finding an optimal group sequential boundary, subject to suitably defined frequentist error probability constraints. This is achieved by first developing a method to solve the unconstrained Bayes problem, with fixed costs for making the wrong decision about

the treatment effect. When a method to solve the unconstrained problem is available, it remains to perform a numerical search for the costs that give a solution with the desired error probabilities.

The problem in Chapter 5 is different in several ways. In Chapter 5, the type I error probability of the phase III trial is not defined by a group sequential boundary that is a part of the optimisation process. The threshold that defines the go/no go decision rule can be thought of as a futility boundary in a one-sided group sequential test, but does not impact the type I error probability of the phase III trial. Moreover, the general formulation in Chapter 5 does not require achieving a certain power at a stipulated effect size. When the problem is re-defined in this way in Section 5.3, the phase III sample size is constrained to take a certain value and there is one less parameter to optimise.

The computational methods that have been used also differ between the chapters. In Chapter 2 the method of dynamic programming can be used to solve the unconstrained decision problem, with fixed costs for the type I and type II error probabilities. The optimisation problem in Chapter 4 was more challenging computationally, as dynamic programming could not be directly applied. The problem could be solved by extending the method of dynamic programming, as described in detail in Section 4.6.

We did not apply the standard method of dynamic programming, illustrated in Chapter 2, to solve the optimisation problem in Chapter 5. There were however still common features with the problems in the other chapters. When the phase III sample size is allowed to depend on phase II data, the problem in Chapter 5 bears similarities with finding the adaptive group sequential designs in Section 2. When n_3 is not allowed to depend on phase II data, there are also connections to the problem considered by Eales and Jennison (1992), who optimise the group sizes of an optimal group sequential test.

6.4 Extensions and future work

Chapter 5 can be viewed as an extension of the other chapters. It brings together methods that we have used for the design of individual trials, to solve the more complex problem of how to design a clinical development programme. As a clinical development program involves many complex decisions and parameters, it is easy to see that the model could be further extended. While some possible extensions were discussed at the end of Chapter 5, the problem of finding the right dose or set of doses to bring forward to phase III would be another very natural extension. In many indications, dose-finding plays a crucial role in the trial immediately before phase III. It would be useful to extend our model by including multiple doses of the experimental drug, as well as a control, in phase II. A very similar framework to the one used in Chapter 5

could be applied, with the key difference that it would be computationally challenging to calculate the posterior distribution of the treatment effect at different dose levels. The optimal phase II sample size is then likely to increase, as more information in phase II would make it easier to identify which dose(s) to bring forward to phase III. It would appear to be reasonable to use a parametric model that assumes a treatment effect that is monotonically increasing in dose. If monotonicity is assumed, it would however also be necessary to model how the toxicity of the drug depends on the dose. Otherwise, any sensible decision rule would end up always selecting the top dose, with the highest efficacy.

6.5 Final words

To summarise, this thesis presents four problems related to the design of clinical trials. It is natural to ask what can be gained by using the methods and designs presented in this thesis. We would argue that the four problems that have been solved are very relevant to clinical trial applications. The solutions that we present are either new approaches to problems that have not yet been solved, or methods that are more efficient than the ones currently available in the literature. If the methods are applied in a correct way, they could ideally contribute to improved clinical trial designs and development programmes. Taking into account the costs of drug development estimated by DiMasi et al. (2003), even modest efficiency gains on a relative scale can translate to important improvements in absolute numbers.

The error spending versions of the optimal non-adaptive group sequential designs in Chapter 2 and Chapter 4 should be easy to implement in practice. They can deal with unpredictable group sizes and could be monitored within a group sequential framework, where a DMC makes recommendations about whether to stop or continue. If desirable, the error spending designs in Chapter 2 could be made non-binding, following the approach used for one-sided tests in Chapter 4. Chapter 3 is also closely related to applications, and what can go wrong if the conditional type I error is not well defined in all situations when a re-design can occur. Finally, the problem in Chapter 5 is probably even closer to clinical trial applications than the others. A model that incorporates what is required for regulatory approval is defined. Thereafter the problem is solved using explicit assumptions about the gains and costs in different phases of the development program. As discussed in Section 6.4, it would be very appealing to make further extensions to the model, as well as applying it to practical examples in different therapeutic areas.

BIBLIOGRAPHY

Ades, A. E., Lu, G., and Claxton, K. (2004). Expected value of sample information calculations in medical decision modeling. *Medical Decision Making*, 24:207–227.

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association*, 58:365–383.

Antonijevic, Z., Pinheiro, J., Fardipour, P., and Lewis, R. J. (2010). Impact of dose selection strategies used in phase II on the probability of success in phase III. *Statistics in Biopharmaceutical Research*, 2:469–486.

Barber, S. and Jennison, C. (2002). Optimal assymetric one-sided group sequential tests. *Biometrika*, 89:49–60.

Bather, J. (2000). *An Introduction to Dynamic Programming and Optimal Stopping*. Wiley.

Bauer, P. (2003). Statistical methodology relevant to the overall development program. *Drug Information Journal*, 37:81–89.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*, 18:1833–1848.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analysis. *Biometrics*, 50:1029–1041.

Bauer, P. and Posch, M. (2004). Letter to the editor. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections by H. Schäfer and H.-H. Müller. Statistics in Medicine 2001; 20:3741-3751. *Statistics in Medicine*, 23:1333–1334.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Brannath, W., Bauer, P., Maurer, W., and Posch, M. (2003). Sequential tests for noninferiority and superiority. *Biometrics*, 59:106–114.

Brannath, W., Bauer, P., and Posch, M. (2006). On the efficiency of adaptive designs for flexible interim decisions in clinical trials. *Journal of Statistical Planning and Inference*, 136:1956–1961.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association*, 97:236–244.

Bretz, F., Schmidli, H., Köning, F., Racine, A., and Maurer, W. (2006). Confirmatory seamless phase II/III trials with hypothesis selection at interim: General concepts (with discussion). *Biometrical Journal*, 48:623–634.

Burman, C.-F., Grieve, A., and Senn, S. (2007). *Decision Analysis in Drug Development.* In Dimitrienko, A., Chuang-Stein, C., and D'Agostino, R. (Eds.), Pharmaceutical Statistics Using SAS: A Practical Guide, Chapter 14. SAS Institute.

Burman, C.-F. and Lisovskaja, V. (2010). The dual test: Safeguarding p-value combination tests for adaptive designs. *Statistics in Medicine*, 29:797–807.

Burman, C.-F. and Senn, S. (2003). Examples of option values in drug development. *Pharmaceutical Statistics*, 2:113–125.

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics*, 62:664–683.

Chandler, G. A. and Graham, I. G. (1988). The convergence of Nystrom methods for Wiener-Hopf equations. *Numerische Mathematik*, 52:345–364.

Chang, W. H. and Chuang-Stein, C. (2004). Type I error and power in trials with one interim analysis. *Pharmaceutical Statistics*, 3:51–59.

Cui, L., Hung, H. M. J., and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, 55:853–857.

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine*, 20:2645–2660.

DiMasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22:151–185.

Dragalin, V. (2006). Adaptive designs: Terminology and classification. *Drug Information Journal*, 40:425–435.

Eales, J. D. and Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika*, 79:13–24.

Eales, J. D. and Jennison, C. (1995). Optimal two-sided group sequential tests. *Sequential Analysis*, 14:273–286.

East-5 (2007). *Software for the design and analysis of flexible clinical trials*. Cytel Software Corporation.

Ellenberg, S. S., Fleming, T. R., and DeMets, D. L. (2003). *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. Wiley.

EMEA (2002). Note for guidance on clinical investigation of medicinal products of diabetes mellitus. Available at `http://www.ema.europa.eu/pdfs/human/ewp/108000en.pdf`. Retrieved on May 26, 2011.

EMEA (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. Available at `http://www.ema.europa.eu/docs/en-GB/documentlibrary/Scientific-guideline/2009/09/WC500003616.pdf`. Retrieved on May 26, 2011.

FDA (2004). Innovation/stagnation: Challenge and opportunity on the critical path to new medical products. Available at `http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html`. Retrieved on May 26, 2011.

FDA (2006). Guidance for clinical trial sponsors: Establishment and operation of clinical trial data monitoring comittees. Available at `http://www.fda.gov.downloads/regulatory/information/guidances/ucm127073.pdf`. Retrieved on May 26, 2011.

FDA (2008). Guidance for industry: Diabetes mellitus - evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. Available at `http://www.fda.gov/downloads/drugs/regulatory/information/guidances/ucm071627.pdf`. Retrieved on May 26, 2011.

FDA (2010). Guidance for industry: Adaptive design clinical trials for drugs and biologics. Available at `http://www.fda.gov/downloads/drugs/regulatory/information/guidances/ucm201790.pdf`. Retrieved on May 26, 2011.

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine*, 17:1551–1562.

Fleming, T. R. and DeMets, D. L. (1996). Surrogate end points in clinical trials: Are we being misled? *Annals of Internal Medicine*, 125:605–613.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall /CRC.

Gittins, J. and Pezeshk, H. (2000a). A behavourial Bayes method for determining the size of a clinical trial. *Drug Information Journal*, 34:355–363.

Gittins, J. and Pezeshk, H. (2000b). How large should a clinical trial be? *The Statistician*, 49:177–187.

Göke, B., Gause-Nilsson, I., and Persson, A. (2007). The effect of tesaglitazar as add-on treatment to metformin in patients with poorly controlled type 2 diabetes. *Diabetes and Vascular Disease Research*, 4(3):204–213.

Gould, A. L. (1997). 3-decision rules for assessing superiority or equivalence. *Proceedings of the ENAR Region of the International Biometrics Society, Memphis, Tennessee*.

Gould, A. L. and Pecore, V. J. (1982). Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika*, 69:75–80.

Hampson, L. V. (2009). *Group Sequential Tests for Delayed Responses*. PhD Thesis, University of Bath.

Home, P. D., Jones, N. P., Pocock, S. J., Beck-Nielsen, H., Gomis, R., Hanefeld, M., Komajda, M., and Curtis, P. (2007). Rosiglitazone RECORD study: glucose control outcomes at 18 months. *Diabetic Medicine*, 24(6):626–634.

ICH (1996). E6 Guideline for good clinical practice. Available at `http://www.ich.org/LOB/media/MEDIA482.pdf`. Retrieved on May 26, 2011.

ICH (1998). E9 Statistical principles for clinical trials. Available at `http://www.ich.org/LOB/media/MEDIA485.pdf`. Retrieved on May 26, 2011.

Inoue, L. Y. T., Thall, P. F., and Berry, D. A. (2002). Seamlessly expanding a randomised phase II clinical trial to phase III. *Biometrics*, 58:823–831.

Jenkins, M., Stone, A., and Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *To appear in Pharmaceutical Statistics*.

Jennison, C. and Turnbull, B. W. (1997). Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92:1330–1341.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall /CRC.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification based on the observed treatment effect. *Statistics in Medicine*, 22:971–993.

Jennison, C. and Turnbull, B. W. (2006a). Adaptive and non-adaptive group sequential tests. *Biometrika*, 93:1–21.

Jennison, C. and Turnbull, B. W. (2006b). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, 25:917–932.

Julious, S. A. and Swank, D. J. (2005). Moving statistics beyond the individual clinical trial: Applying decision science to optimize a clinical development plan. *Journal of the American Statistical Association*, 94:468–482.

Koch, A. (2006). Confirmatory clinical trials with an adaptive design. *Biometrical Journal*, 48:574–585.

Koyama, K., Sampson, A. R., and Gleser, L. J. (2005). A framework for two-stage adaptive procedures to simultaneously test non-inferiority and superiority. *Statistics in Medicine*, 24:2439–2356.

Lachin, J. M. (2005). A review of methods for stopping based on conditional power. *Statistics in Medicine*, 24:2747–2764.

Lai, T. (1973). Optimal stopping and sequential tests which minimise the maximum sample size. *Annals of Statistics*, 1:659–673.

Lai, T., Shih, M.-C., and Zhu, G. R. (2006). Modified Haybittle-Peto group sequential designs for testing superiority and non-inferiority hypotheses in clinical trials. *Statistics in Medicine*, 25:1149–1167.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663.

Lan, K. K. G., Lachin, J. M., and Bautista, O. (2003). Over-ruling a group sequential boundary - a stopping rule versus a guideline. *Statistics in Medicine*, 22:3347–3355.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55:1286–1290.

Lesko, L. J. and Atkinson, A. J. (2001). Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: Criteria, validation, strategies. *Annual Review Pharmacological Toxicology*, 41:347–366.

Lindley, D. V. (1997). The choice of sample size. *The Statistician*, 46:129–138.

Liu, Q., Anderson, K. M., and Pledger, G. W. (2004). Benefit-risk evaluation of multi-stage adaptive designs. *Sequential Analysis*, 23:317–331.

Martinez-Torres, F., Menon, S., Pritsch, M., Victor, N., Jenetzky, E., Jensen, K., Schielke, E., Schmutzhard, E., de Gans, J., Chung, C., Luntz, S., Hacke, W., and Meyding Lamadé, U. (2008). Protocol for German trial of acyclovir and corticosteroids in herpes-simplex-virus-encephalitis (GACHE): a multicenter,

multinational, randomized, double-blind, placebo-controlled, German, Austrian and Dutch trial. *BMC Neurology*, 8:40.

Mehta, C. R. and Tsiatis, A. A. (2001). Flexible sample size considerations using information based interim monitoring. *Drug Information Journal*, 35:1095–1112.

Morikawa, T. and Yoshida, M. (1995). A useful testing strategy in phase III trials: Combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics*, 5:297–306.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential designs. *Biometrics*, 57:886–891.

Müller, H.-H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine*, 23:2497–2508.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7:308–313.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley.

O'Hagan, A., Stevens, J. W., and Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, 4:187–201.

Öhrn, F. and Jennison, C. (2010). Optimal group sequential designs for simultaneous testing of superiority and non-inferiority. *Statistics in Medicine*, 29:743–759.

Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. *Journal of Statistical Planning and Inference*, 42:19–35.

Pampallona, S., Tsiatis, A. A., and Kim, K. (2001). Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Information Journal*, 35:1113–1121.

Parmigiani, G. and Inoue, L. Y. T. (2009). *Decision Theory - Principles and Approaches*. Wiley.

Pezeshk, H., Nematollahi, N., Maroufy, V., and Gittins, J. (2009). The choice of sample size, a mixed Bayesian/frequentist approach. *Statistical Methods in Medical Research*, 18:183–194.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199.

Posch, M., Timmesfeld, N., König, F., and Müller, H.-H. (2004). Conditional rejection probabilities of student's t-test and design adaptations. *Statistics in Medicine*, 46:389–403.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and opterational criteria. *Statistics in Medicine*, 8:431–440.

Proschan, M. A., Follmann, D. A., and Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 48:1131–1143.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of clinical studies based on conditional power. *Biometrics*, 51:1315–1324.

Proschan, M. A., Lan, K. K. G., and Wittes, J. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer.

Rinke, A., Müller, H.-H., Schade-Brittinger, C., Klose, K.-J., Barth, P., Wied, M., Mayer, C., Aminossadati, B., Pape, U.-F., Bläker, M., Harder, J., Arnold, C., Gress, T., and Arnold, R. (2009). Placebo-controlled, double-blind, prospective, randomized study on the effect of octreotide LAR in the control of tumour growth in patients with metastatic neuroendocrine midgut tumors: A report from the PROMID study group. *American Society of Clinical Oncology*, 27:4656–4663.

Schäfer, H. and Müller, H.-H. (2001). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine*, 20:3741–3751.

Schäfer, H., Timmesfeld, N., and Müller, H.-H. (2006). An overview of statistical approaches for adaptive designs and design modifications. *Biometrical Journal*, 48:507–520.

Scherag, A., Hedebrand, J., Müller, H.-H., and Schäfer, H. (2009). Flexible designs for genom-wide association studies. *Biometrics*, 65:815–821.

Schmidli, H., Bretz, F., Racine, A., and Maurer, W. (2006). Confirmatory seamless phase II/III trials with hypothesis selection at interim: Applications and general considerations. *Biometrical Journal*, 48:635–643.

Schmitz, N. (1993). *Optimal sequentially planned decision procedures*. Springer.

Schoenfeld, D. (1980). Statistical considerations for pilot studies. *International Journal of Radiation Oncology Biology Physics*, 6:371–374.

Senn, S. (1997). *Statistical Issues in Drug Development*. Wiley.

Shih, W. J., Quan, H., and Gang, L. (2004). Two-stage adaptive strategy for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine*, 23:2781–2798.

Snapinn, S. (2006). Assessment of futility in clinical trials. *Pharmaceutical Statistics*, 5:273–281.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, 7:8–17.

Stallard, N. (1998). Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics*, 54:279–294.

Stallard, N. (2003). Decision-theoretic designs for phase II clinical trials allowing for competing studies. *Biometrics*, 59:402–209.

Stallard, N. (2010). A confirmatory seamless phase II/III clinical trial incorporating short-term endpoint information. *Statistics in Medicine*, 29:959–971.

Stallard, N., Thall, P. F., and Whitehead, J. (1999). Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics*, 55:971–977.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine*, 22:689–703.

Stallard, N., Whitehead, J., Todd, S., and Whitehead, A. (2001). Stopping rules for phase II studies. *British Journal of Clinical Pharmacology*, 51:523–529.

Timmesfeld, N., Schäfer, H., and Müller, H.-H. (2006). Increasing the sample size during clinical trials with t-distributed test statistics without inflating the type I error rate. *Statistics in Medicine*, 26:2449–2464.

Todd, S. and Stallard, N. (2005). A new clinical trial design combining phases 2 and 3: Sequential designs with treatment selection and a change of endpoint. *Drug Information Journal*, 39:109–118.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90:367–378.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16:117–186.

Wald, A. (1947). *Sequential Analysis*. Wiley.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19:326–339.

Wang, S. J., Hung, H. M. J., Tsong, Y., and Cui, L. (2001). Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine*, 20:1903–1912.

Wang, S. J., Hung, M. J., and O'Neill, R. T. (2006). Adapting the sample size planning of a phase III trial based on phase II data. *Pharmaceutical Statistics*, 5:85–97.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193–200.

Whitehead, J. (1985). Designing phase II studies in the context of a programme of clinical research. *Biometrics*, 5:373–383.

Whitehead, J. (1986). Sample sizes for phase II and phase III clinical trials: an integrated approach. *Statistics in Medicine*, 5:459–464.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Wiley.

World Medical Association (2008). Declaration of Helsinki - ethical principles for medical research involving human subjects. Available at `http://www.wma.net/en/30publications/10policies/b3/index.html`. Retrieved on May 26, 2011.

Yin, Y. (2002). Sample size calculation for a proof of concept study. *Journal of Biopharmaceutical Statistics*, 12:267–276.